UNIVERSIDADE ESTADUAL DE CAMPINAS - UNICAMP

INSTITUTO DE MATEMÁTICA, ESTATÍSTICA E

CIÊNCIA DA COMPUTAÇÃO - IMECC

DISCRIMINAÇÃO COM MISTURA DE VARIÁVEIS CONTÍNUAS E CATEGÓRICAS

MERCEDES ANA VALDIVIA LEÓN

Profa. Dra. REGINA C. C. P. MORAN
ORIENTADORA

Campinas - SP

Agosto, 1993.

" DISCRIMINAÇÃO COM MISTURA DE VARIÁVEIS CONTÍNUAS E CATEGÓRICAS"

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pela Sra. Mercedes Ana Valdivia León e aprovada pela Comissão Julgadora.

Campinas, 17 de Setembro de 1993

Profa. Dra. Regina C.C.P. Moran

Dissertação apresentada ao Instituto de Matemática, Estatística e Ciência da Computação, UNICAMP, como requisito parcial para obtenção de Título de MESTRE em Estatística.

AGRADECIMENTOS

A Profa. Dra. Regina C.C.P. Moran, pelos conhecimentos transmitidos, apoio e orientação na elaboração deste trabalho.

Ao Conselho Nacional de Desenvolvimento Científico e Tecnológico (CNPq), pela concessão da bolsa de mestrado por 30 meses.

Ao Fundo de Apoio ao Ensino e à Pesquisa (FAEP/UNICAMP), pela contribuição para afinalização deste trabalho através do auxílio ponte.

Ao Dr. W. J. Krzanowski pelo fornecimento do programa computacional para aplicar o método de discriminação baseado no Modelo de Posição.

A Rissa e ao Flávio pela digitação do trabalho e a colaboração.

RESUMO

Neste trabalho trata-se o problema de discriminação entre duas populações caracterizadas por algumas variáveis categóricas e outras contínuas. A presença das variáveis categóricas faz com que, dentro de cada população, apareçam sub-populações que devem ser consideradas na análise.

Primeiramente estuda-se a estrutura particular do vetor de médias e da matriz de variância-covariância de três tipos de vetores aleatórios mistos de interesse. Em seguida, descreve-se alguns métodos de discriminação propostos na literatura, para tratar problemas com misturas. Finalmente apresenta-se um exemplo que permite ilustrar os métodos de interesse, compará-los e mostrar algumas de suas limitações em aplicações práticas.

SUMÁRIO

CAPÍTULO 1

INTRODUÇÃO

1.1 - MOTIVAÇÃO

Na análise multivariada de dados agrupados, quando há informação externa sobre procedências de um número de populações possivelmente distintas, o problema de discriminação é tratado através da Análise Discriminante. Este método permite resolver o seguinte problema de classificação.

Dado um objeto (ou pessoa) que se sabe pertence a algum de g grupos, exclusivos e exaustivos, π_1, ..., π_g em que foi dividida uma população π, trata-se de alocar o objeto a um destes grupos baseado em um conjunto de características medidas nele.

O objetivo fundamental da Análise Discriminante é construir uma regra de alocação que seja ótima em algum sentido tal como minimizar o número esperado de erros de classificação ou o custo médio de

classificar mal futuras observações. Este método também pode ser usado com a finalidade de caracterizar as diferenças entre os grupos ou para determinar um conjunto de variáveis que reflita da "melhor maneira" possível estas diferenças.

A literatura sobre Análise Discriminante é extensa. Isto pode ser visto, por exemplo, na bibliografia de Cacoullos (1973), Lachenbruch (1975), Huberty (1975), Goldstein e Dillon (1978) e Lachenbruch e Goldstein (1979). Este tema às vezes é tratado sob o título de classificação embora este termo usualmente fique reservado para a Análise de Conglomerados. Na literatura sobre reconhecimento de padrões este tipo de problema é conhecido como reconhecimento de padrões estatístico com aprendizagem supervisionado (*statistical pattern recognition with supervised learning*) (ver Dolby, 1970; Duda e Hart,1973).

Existe grande variedade de situações que dão origem a problemas de discriminação, por exemplo:

- Os alunos de certa escola devem fazer, ao final do ano, um exame que determina se passam ou não à série seguinte, isto os separa em dois grupos exclusivos e exaustivos: aprovados e reprovados. Um dos alunos perde o exame por razões de saúde e o professor deve decidir se o aluno passa de série baseado em um conjunto de notas de provas e trabalhos feitos pelo aluno ao longo do ano.
- O gerente de pessoal da empresa A deve contratar 50 novos funcionários. A partir de uma avaliação detalhada dos atuais funcionários da empresa, estes são divididos em três grupos: competentes, regulares e incompetentes. Trata-se de escolher um conjunto de características que identifiquem os funcionários competentes para facilitar a avaliação e contratação do pessoal novo.

No primeiro exemplo, as variáveis usadas para classificar o aluno

que não fez a prova são consideradas contínuas (notas), no segundo exemplo, as características a serem observadas nos candidatos são do tipo categórico. Na prática não é tão frequente encontrar problemas de discriminação onde todas as variáveis envolvidas são do mesmo tipo, é muito comum que os grupos estejam caraterizados por misturas de diferentes tipos de variáveis, desde variáveis contínuas até variáveis categóricas não ordenadas. Por exemplo, em um estudo sobre o solo, isto está caraterizado por medidas contínuas de laboratório como o pH e atributos categóricos como cor e textura; na medicina algumas doenças podem ser identificadas pela ausência ou presença de certos sintomas (variáveis binárias), além do aumento na temperatura ou pressão dos pacientes (variáveis contínuas).

A maior diferença entre as variáveis contínuas e categóricas no contexto de discriminação é que as últimas geram subpopulações nas quais o modelo de discriminação pode ser diferente, por exemplo: homens e mulheres poderiam ter distintos sintomas da mesma doença, ou a variedade A do milho produzir mais que a variedade B se não for atacada por uma certa praga porém, pode produzir menos para algum nível de ataque.

Se as variáveis categóricas envolvidas no problema de discriminação são ordenadas, Krzanowski (1980) sugere atribuir valores numéricos a cada uma das categorias e depois tratar todas as variáveis como se fossem contínuas. Isto leva a usar as técnicas desenvolvidas para o caso contínuo como a Função Linear Discriminante (FLD) introduzida por Fisher em 1936. No caso de variáveis categóricas não ordenadas, muitas vezes os analistas procedem da mesma maneira sem parar para pensar nas consequências que isto pode ter no desempenho da técnica ou nos resultados obtidos a partir dela. Como alternativa Cochran e Hopkins (1961) propuseram categorizar as variáveis contínuas e depois usar técnicas para variáveis discretas (ver Goldstein e Dillon, 1978). Krzanowski (1980) comenta que outra solução possível é

separar os dois tipos de variáveis para efetuar duas análises separadas usando técnicas apropriadas para cada caso.

Nenhuma das soluções anteriores é a mais adequada. Tratar uma variável nominal como contínua significa atribuir-lhe um nível de medida superior ao verdadeiro, isto é perigoso pois equivale a aceitar a validade de noções de ordem e/ou razão entre os valores numéricos atribuídos às categorias. Por outro lado, converter uma variável contínua em ordinal ou nominal resulta em perda de informação que poderia ser útil aos objetivos da pesquisa.

Os problemas acima descritos têm motivado a criação de novas técnicas, ou a modificação de algumas já existentes, para resolver de maneira ótima os problemas de discriminação em presença de mistura de variáveis.

1.2 - O PROBLEMA E COMO É TRATADO

A Análise Discriminante foi descrita na seção anterior como um método apropriado para classificar observações em g ($g \geq 2$) grupos porém, em muitas aplicações práticas $g=2$ ou seja, o problema se reduz a alocar observações em dois grupos ou a descrever as diferenças entre dois grupos. Por exemplo, pacientes normais ou não, alunos reprovados ou aprovados, ossos de fêmeas ou machos e notas verdadeiras ou falsas.

Fazendo uma revisão das publicações sobre Análise Discriminante encontra-se que a maior parte dos trabalhos tratam assuntos relacionados ao caso de dois grupos, as extensões para três ou mais grupos usualmente aparecem como generalizações de estudos feitos anteriormente para aquele caso particular. Na literatura sobre discriminação com mistura de variáveis o padrão anterior se repete. Este fato contribuiu na definição da abrangência deste trabalho e do

problema específico que será aqui estudado e que é descrito a seguir.

Tratar-se-á de discriminar entre dois grupos (ou populações) π_1 e π_2, baseado num vetor misto $\mathbf{w}'=(\mathbf{x}',\mathbf{y}')$ composto por q variáveis categóricas $x_1, \ldots, x_q$ arranjadas no vetor $\mathbf{x}$ e p variáveis contínuas $y_1, \ldots, y_p$ arranjadas no vetor $\mathbf{y}$.

A construção da regra de classificação depende, como em qualquer problema de discriminação do método de discriminação escolhido porém neste caso a escolha do método depende principalmente da quantidade de dados disponível, ou seja, do tamanho das amostras de treinamento (*training samples*).

Seguindo a classificação de Seber (1984), os métodos de discriminação para tratar misturas propostos na literatura, podem ser agrupados em quatro classes levando em conta a informação disponível sobre os grupos. Nestas classes os grupos são: (1) completamente conhecidos, (2) conhecidos exceto por alguns parâmetros desconhecidos, (3) parcialmente conhecidos e (4) completamente desconhecidos. Seber (1984) afirma que na prática, a diferença entre (1) e (2) está no número de observações que compõem as amostras de cada grupo, já que para amostras muito grandes a variação nas estimativas dos parâmetros pode ser ignorada.

Como representante da classe (1) considere-se o método de discriminação baseado no Modelo de Posição (tradução que será usada neste trabalho para *Location Model*) , proposto por Krzanowski (1975,1980). A característica principal deste método é que o autor segue um enfoque estritamente paramétrico e como primeiro passo busca um modelo adequado para a distribuição do vetor misto $\mathbf{w}$. O modelo escolhido foi o modelo de posição sugerido por Olkin e Tate (1961) que assume que as variáveis categóricas representam categorias multinimiais e que a distribuição condicional das variáveis contínuas,

dadas as categóricas, é normal multivariada com vetor de médias que depende da população e da categoria observada, e matriz de variância covariância comum a todas as categorias e a mesma nas duas populações. Já com o modelo adequado para o vetor misto **w**, o passo seguinte foi construir uma regra de alocação pelo princípio da razão de verossimilhança que resultou em um conjunto de hiperplanos de discriminação, um para cada categoria multinomial.

Para representar a classe (2) basta considerar o caso amostral do método anterior onde os parâmetros são desconhecidos e devem ser estimados.

Entre os métodos da classe (3), que supõe populações parcialmente conhecidas, podem ser mencionadas a FLD de Fisher (adaptada para misturas por Krzanowski, 1975) e as três Funções Lineares Discriminantes Modificadas propostas por Vlachonikolis e Marriott (1982) que incorporam, como variáveis independentes, novas variáveis construidas a partir das variáveis observadas. Isto, com a intenção de representar na função de discriminação, as interações entre variáveis categóricas e entre variáveis categóricas e contínuas que têm caráter mais multiplicativa do que aditiva e que por tanto não são devidamente representadas por uma função linear das variáveis originais. Estes métodos são considerados nesta classe, embora não sejam derivados fazendo suposições sobre a distribuição das variáveis nas populações, pois a construção das funções de discriminação depende dos vetores de médias e da matriz de variância covariância (suposta comum) das variáveis nos grupos, portanto estes parâmetros devem ser conhecidos ou estimados.

Ainda na classe (3), tem-se o método de Discriminação Logística proposto primeiramente em Cornfield (1962) e depois em Cox (1966) e Day e Kerridge (1967). A suposição básica deste método é que o logarítmo da razão das verossimilhanças é linear nas variáveis, isto

leva à construção de uma regra de classificação baseada nas probabilidades a posteriori que com ajuda do Teorema de Bayes são escritas, de maneira simples e fácil de calcular, dependendo unicamente de uma função linear nas variáveis cujos coeficientes são estimados usando procedimentos de máxima verossimilhança. Um detalhe interessante relacionado à Discriminação Logística é que alguns pesquisadores como Press e Wilson (1978) e Knoke (1982) não a consideram um método propriamente dito senão uma maneira alternativa de estimar os coeficientes das funções lineares de discriminação.

Finalmente, na classe (4) tem-se os métodos que supoem populações totalmente desconhecidas. De um lado o enfoque baseado na estimação Kernel com uma proposta específica para dados mistos de Aitchinson e Aitken (1976) e de outro lado a classificação segundo o vizinho mais próximo baseada nos métodos propostos por Cover e Hart (1967) e adaptado para misturas por Wojciechowski (1987).

1.3 - ALGUMAS EXTENSÕES DO PROBLEMA

Na seção 1.2 foi definido o problema específico de discriminação em presença de mistura de variáveis que será tratado neste trabalho. Ainda, foram classificados e descritos brevemente alguns procedimentos de classificação apropriados que serão estudados com mais detalhe nos capítulos seguintes. Nesta seção são apresentados alguns desenvolvimentos e extensões dos métodos, com a intenção de completar uma visão geral sobre o tratamento que tem recebido a discriminação com mistura de variáveis na literatura.

A proposta que originou o maior número de extensões é a baseada no modelo de posição. Além dos dois artigos básicos onde são apresentados o procedimento de classificação para mistura de variáveis binárias e contínuas (Krzanowski, 1975) e a extensão do método para

misturas de variáveis contínuas e categóricas em geral (Krzanowski, 1980), o autor da proposta, Krzanowski, tem contribuído com vários outros trabalhos relacionados ao método que serão mencionados a seguir, em ordem cronológica de publicação.

Em 1976, Krzanowski sugere estudar a eficiência do modelo de posição através da representação gráfica das distâncias de Mahalanobis entre os grupos, via *principal coordinate analysis* (ver Gower, 1966 e Krzanowski, 1971). Este artigo mostra como a aplicação do modelo de posição a um conjunto de dados pode ser examinada de tal forma a revelar a estrutura dos dados e talvez apontar a razão pela qual qualquer melhora de desempenho tenha sido observada em relação à FLD de Fisher.

Em 1979, propõe algumas transformações lineares adequadas para observações mistas que simplificam sua estrutura e facilitam as análises preliminares dos dados. Com relação à Análise Discriminante, as transformações reduzem a dimensão do problema produzindo novas variáveis em menor número que as originais, e com uma estrutura adequada, que permite o uso da FLD de Fisher em lugar de procedimentos mais complicados que seriam os indicados para as variáveis originais. Por último, sugere um teste de ajuste para a FLD baseada nas variáveis transformadas.

No ano 1982, Krzanowski extende, para o caso de misturas de variáveis, o enfoque baseado em testes de hipóteses para discriminar duas populações multinormais de Anderson (1958) e Jhon (1960, 1963). O autor deriva uma regra de alocação baseada no modelo de posição usando argumentos de testes de hipóteses e ainda discute brevemente, alguns dos possíveis méritos do novo enfoque.

Krzanowski (1983a) propõe um método *backward* de seleção de variáveis categóricas que permite identificar um modelo de posição

reduzido e adequado para aplicações em discriminação, quando o número de variáveis categóricas é muito grande, e todas elas não podem ser usadas diretamente. No mesmo ano, Krzanowski (1983b) deriva uma medida de distância entre populações caracterizadas por uma forma geral do modelo de posição e discute a estimação dos parâmetros para essa medida. Este artigo constituiu a base da extensão do método de discriminação baseado no modelo de posição para classificar mais de duas populações proposto em Krzanowski (1986). Nesse artigo mostra-se que para o caso de dois grupos, a classificação por mínima distância é equivalente à regra de classificação de máxima verossimilhança baseada no modelo de posição, isto possibilita a extensão para discriminação múltipla (no sentido do Krzanowski, 1986, ou seja, de mais de dois grupos). Um detalhe importante a ser colocado é que se for necessário considerar probabilidades prévias ou custos de classificação errada, então os procedimentos mais complexos de comparação de funções de discriminação duas a duas devem ser usados. Como alternativa neste contexto, o autor sugere basear a classificação multipla nas FLD Modificadas propostas por Vlachonikolis e Marriott (1982), estimando os coeficientes por regressão logística.

Outro pesquisador interessado nos temas relacionados à discriminação baseada no modelo de posição é I. Vlachonikolis. Para começar, pode ser considerado o trabalho publicado em 1982 feito com a colaboração de F.H.C. Marriott no qual são comparadas algumas técnicas de discriminação para mistura de variáveis (entre elas a baseada no modelo de posição). Já em 1985, Vlachonikolis apresenta uma extensão assintótica da distribuição da regra de alocação baseada no modelo de posição para o caso que os parâmetros são substituidos pelos estimadores amostrais comuns. No seu trabalho seguinte Vlachonikolis (1986) compara as estimativas da probabilidade esperada de classificação errada, quando o modelo de posição estimado é usado, obtidas a partir da aproximação baseada na expansão assintótica derivada em 1985, com as estimativas obtidas por métodos Monte Carlo.

Os resultados apresentados para diversas combinações de tamanhos de amostra, número de variáveis binárias e contínuas, distâncias de Mahalanobis e probabilidades multinomiais, suportam a conclusão de que a aproximação é boa inclusive para tamanhos de amostra moderados.

Leung (1989) também trabalhou na distribuição assintótica da função de discriminação estimada baseada no modelo de posição. Ele considerou a função padronizada, derivando a sua distribuição assintótica, sem precisar inverter a função característica correspondente ou conhecer os valores exatos das distâncias de Mahalanobis que eram necessários para derivar a expansão de Vlachonikolis (1985).

Para terminar com as contribuições de Vlachonikolis neste tema, tem-se o artigo publicado em 1990 no qual o autor deriva uma regra de classificação preditiva para misturas baseando este enfoque nas distribuições de frequências usuais associadas ao modelo de posição e distribuições a priori vagas para os parâmetros desconhecidos. Apresenta também, alguns resultados Monte Carlo para comparar o desempenho da nova regra preditiva com a regra estimada concluindo que, sob as condições consideradas, o valor esperado das taxas de erro se comporta em geral, de maneira análoga para os dois métodos.

Finalmente, tem-se o trabalho de Tu e Han (1982) no qual é proposto um esquema de amostragem chamado *Double Inverse Sampling* que garante que as matrizes de covâriancia estimadas não sejam singulares. Este esquema é derivado para o caso particular do modelo de posição onde são consideradas p variáveis contínuas e uma binária (a regra de classificação para este modelo foi derivada por Chang e Afifi, 1974).

Outra proposta que motivou algumas extensões interessantes é a Discriminação Logística. Anderson (1982) menciona e comenta as extensões aqui descritas brevemente. Para maiores detalhes

recomenda-se consultar esse artigo e as referências nele.

Os temas desenvolvidos nas extensões são:

i) Seleção de variáveis : Anderson (1982) sugere um método *stepwise* para escolher um conjunto de boas variáveis preditoras (dentre as possíveis), apropriado para a Discriminação Logística. Neste método o critério para selecionar modelos é comparar o valor máximo do logarítmo da função de verossimilhança obtido com cada modelo. Por outro lado, a variável candidata a ingressar no modelo, em cada passo do método, é avaliada com uma estatistica com distribuição assintótica chi-quadrado.

ii) Discriminação Logística Quadrática : Anderson (1975) considerou o caso em que o logarítmo da razão das verossimilhanças é uma função quadrática do vetor de variáveis e sugeriu uma aproximação da forma quadrática envolvida, para o caso em que o número de parâmetros cresce muito e não é possível usar os procedimentos de estimação iterativos.

iii) Redução do vício na Discriminação Logística : O fato dos estimadores de máxima verossimilhança baseados em amostras pequenas, serem viciados é bem conhecido. Cox e Hinkley (1974) propuseram algumas técnicas de correção que permitem eliminar parcialmente este vício. Anderson e Richardson (1979) consideraram uma aplicação desses métodos para encontrar estimadores corrigidos dos coeficientes da função de discriminação logística, cujas propriedades foram avaliadas usando simulação. A conclusão foi que o melhor desempenho dos estimadores propostos, em relação aos estimadores de máxima verossimilhança, é mais notório quando as amostras de treinamento não são muito pequenas.

iv) Atualização da função de discriminação (*Discriminant updates*). Anderson (1979) sugere aproveitar a informação sobre os parâmetros logísticos contida em observações que ainda não foram

alocadas em nenhuma população. Para isto ele construiu uma nova função de verossimilhança que considera as observações do grupo 1, grupo 2 e do grupo 3, o último das observações não classificadas, cuja distribuição é uma mistura das distribuições dos grupos 1 e 2 com proporções θ e $(1-\theta)$ respectivamente. O custo deste procedimento é a introdução de um parâmetro adicional a ser estimado. Anderson (1982) comenta que este tipo de extensão não existe para a FLD mas que Murray e Titterington (1978) abordaram este problema usando métodos de discriminação kernel.

v) Discriminação Logística para mais de dois grupos : A extensão do método para discriminar três ou mais grupos é direta. Assume-se que o logarítmo da razão das verossimilhanças é linear para qualquer par de populações. Depois estima-se os parâmetros usando métodos de máxima verossimilhança e por último as novas observações são alocadas na população com maior probabilidade posterior. Todas as extensões da Discriminação Logística mencionadas antes podem ser adaptadas para Discriminação Logística multipla, a única restrição está dada pelo número de parâmetros que deve ser estimado e que deve ficar dentro dos limites operacionais dos procedimentos de otimização.

Antes de terminar esta seção é importante comentar que não foram encontradas extensões particulares para as FLD Modificadas propostas por Vlachonikolis e Marriott (1982). Uma razão possível para isto é que seu tratamento teórico e computacional é totalmente análogo ao da FLD de Fisher portanto, as extensões propostas para a última são válidas para as modificações (isto inclui métodos de seleção de variáveis, métodos para calcular taxas de erro e também cálculo dos coeficientes das funções por regressão logística). A maior vantagem disto é que não é necessário novo *software*, qualquer programa computacional comum para Análise Discriminante pode ser usado.

1.4 - SOBRE O TRABALHO

Este trabalho tem três objetivos fundamentais que são expostos a seguir:

- Chamar a atenção sobre a necessidade de um tratamento adequado para os problemas nos quais algumas variáveis envolvidas são contínuas e outras categóricas, mostrando a estrutura particular das observações mistas e as inter-relações entre os dois tipos de variáveis.
- Apresentar e descrever alguns métodos desenvolvidos na literatura para resolver problemas de discriminação na presença de misturas de variáveis contínuas e categóricas e ainda, mostrar os programas computacionais disponíveis para aplicar estes métodos.
- Apresentar alguns resultados comparativos sobre os métodos, que facilitem a escolha de um método adequado para discriminar com mistura de variáveis, levando em conta a informação disponível sobre as populações e os tamanhos das amostras de treinamento.

Baseado nestes objetivos, o trabalho foi organizado da seguinte maneira.

No capítulo 2 apresenta-se as definições e transformações necessárias para o desenvolvimento posterior dos métodos de discriminação. Ainda, deriva-se o vetor de médias e a matriz de variância-covariância, populacionais e amostrais, de três tipos de vetores aleatórios mistos de interesse. A análise do primeiro e segundo momentos mostra características particulares das observações mistas e revela as componentes discretas da mistura como fontes de variação, pois impõem uma sobre-estrutura de grupos nas populações.

No capítulo 3 são descritos em detalhe os seguintes métodos de discriminação para misturas de variáveis contínuas e categóricas :

- Método baseado no modelo de posição
- FLD de Fisher adaptada para misturas
- Funções Lineares Discriminantes Modificadas.

Os métodos acima constituem os principais focos de interesse deste trabalho no entanto, não são os únicos considerados aqui, tem-se também a Discriminação Logística descrita de maneira ampla e outros dois métodos: a Discriminação Kernel e a Discriminação Segundo o Vizinho mais Próximo que são tratados brevemente.

No capítulo 4 apresenta-se uma aplicação prática para ilustrar os métodos de interesse. O conjunto de dados utilizado e algumas análises iniciais são apresentados no Apêndice A.

Para terminar, no capítulo 5 são expostas algumas considerações finais e as conclusões do trabalho.

CAPÍTULO 2

CARACTERIZAÇÃO BÁSICA DE MISTURAS

Neste capítulo é feita uma caracterização do vetor aleatório misto no sentido de definições, formas equivalentes de tratá-lo via transformações e derivação do vetor de médias e matriz de variância-covariância, por estes motivos chamada caracterização básica.

Os elementos tratados neste capítulo apareceram primeiramente como necessários para o desenvolvimento dos métodos de discriminação descritos no capítulo 3.

A seção 2.1 está dividida em duas sub-seções. Na sub-seção 2.1.1 são apresentadas as definições essenciais para desenvolvimentos posteriores. Na sub-seção 2.1.2 são tratadas transformações de variáveis categóricas em binárias e/ou multinomiais. Estas transformações são uma compilação daquelas encontradas nos distintos métodos a serem tratados no capítulo 3. São de fato, maneiras

alternativas de olhar o mesmo problema e seus objetivos principais são,

i) Mostrar que não há perda de generalidade quando se estuda mistura de variáveis binárias e contínuas , pois o caso geral, de misturas de variáveis categóricas e variáveis contínuas, sempre pode ser reduzido àquele.

ii) Representar cada arranjo do vetor de variáveis categóricas como uma categoria multinomial. Isto permite estudar individualmente cada um desses arranjos e as suas interações com as variáveis contínuas na mistura.

Na seção 2.2 serão calculadas as esperanças e as matrizes de variância-covariância populacionais e amostrais de três vetores mistos: $\mathbf{w}$ composto por q variáveis binárias e p contínuas, $\mathbf{w}^*$ composto por uma multinomial e p contínuas e $\mathbf{v}$ composto por uma multinomial, p contínuas e os produtos cruzados dos elementos da multinomial com as variáveis contínuas. O vetor $\mathbf{v}$ é de interesse pois será usado como base para construir uma FLD modificada na seção 3.4.

A caracterização e interpretação dos momentos amostrais básicos, médias, variâncias e covariâncias é uma contribuição deste trabalho. As expressões algébricas encontradas por adequada formulação da matriz de dados exibem os esperados estimadores de momentos das correspondentes expressões populacionais. Estas por sua vez revelam com muita clareza o efeito do tratamento conjunto das variáveis na presença da mistura.

Os resultados destes procedimentos algébricos mostram a estrutura da observação mista e as interrelações entre variáveis categóricas e contínuas pelo simples exame do vetor de médias e matrizes de variância-covariância.

Os momentos das variáveis contínuas na presença das categorias, mostram estas últimas como fontes de variação atuando como sobrestrutura de grupos sobre as populações ou indivíduos. De fato, o vetor de médias das contínuas tem como componentes médias ponderadas das médias "dentro" de categorias. As matrizes de variância e covariância das contínuas exibem no cálculo das variâncias e covariâncias a contribuição da variabilidade dentro das categorias e entre categorias. Além disso, as covariâncias entre categóricas e contínuas refletem o efeito da categoria na média: covariância entre categórica e contínua se reduz a uma média ponderada, pelas probabilidades de cada categoria, dos efeitos da categoria na média, ou seja, desvio entre a média de cada contínua dentro da categoria e uma média geral.

Estes resultados são úteis na análise de misturas e sugerem uma apresentação da matriz de variância e covariância que permita recuperar a participação de cada fonte e/ou peso nos primeiro ou segundo momentos sob análise.

2.1 - DEFINIÇÕES E TRANSFORMAÇÕES

2.1.1 - Definições Básicas

Definição 2.1.1 Um vetor aleatório $\mathbf{w} = (w_1, w_2, \ldots, w_{p+q})'$ é dito um vetor aleatório misto se q (q > 0) das variáveis aleatórias que o compõem são discretas e as outras p (p > 0), são contínuas. Isto é, o vetor $\mathbf{w}$ pode ser escrito como $\mathbf{w} = (\mathbf{w}'_{(1)}, \mathbf{w}'_{(2)})'$ onde:

$\mathbf{w}_{(1)}$ é um vetor discreto q - variado e

$\mathbf{w}_{(2)}$ é um vetor contínuo p - variado.

Neste trabalho se assumirá que as componentes discretas do vetor misto são categóricas, portanto, serão consideradas unicamente misturas de variáveis categóricas (ordinais e nominais) e variáveis contínuas.

Definição 2.1.2 Uma variável aleatória x, é dita variável aleatória binária se assume unicamente os valores zero e um.

Definição 2.1.3 Distribuição Multinomial (Johnson e Kotz, 1969 p. 281). Considere uma série de ensaios independentes nos quais somente um de k eventos mutuamente exclusivos e exaustivos $E_1, \ldots, E_k$ deve ser observado e tais que a probabilidade de ocorrência do evento E_j é p_j para cada ensaio (com $\sum_{j=1}^{k} p_j = 1$). Então a distribuição conjunta do vetor aleatório $(n_1, \ldots, n_k)'$ cujas componentes representam os números de ocorrência dos eventos $E_1, \ldots, E_k$, respectivamente, em N ensaios (com $\sum_{j=1}^{k} n_j = N$) está definida por:

$$P(n_1, n_2, \ldots, n_k) = N! \prod_{j=1}^{k} \left(\frac{p_j^{n_j}}{n_j!} \right) \qquad \left(0 \leq n_j ; \sum_{j=1}^{k} n_j = N \right)$$

e é usualmente chamada distribuição multinomial com parâmetros N, $p_1, p_2, \ldots, p_k$ e denotada Multin(N, $p_1, \ldots, p_k$).

Propriedade 2.1.1 A distribuição conjunta de qualquer subconjunto $n_{a1}, \ldots, n_{as}$ das n_j's também é multinomial (com uma variável (s+1) igual a $N - \sum_{j=1}^{s} n_{a_j}$). De fato:

$$P(n_{a_1},\ldots,n_{a_s}) = N![(N-\sum_{j=1}^{s} n_{a_j})! \prod_{j=1}^{s} n_{a_j}!]\ (1-\sum_{j=1}^{s} p_{a_j})^{N-\sum_{j=1}^{s} n_{a_j}} \prod_{j=1}^{s} p_{a_j}^{n_{a_j}}$$

Observe-se que a distribuição binomial b(N, p_1) cumpre a propriedade anterior fazendo s = 1.

2.1.2 - Transformação de Variáveis Categóricas

A seguir serão definidas algumas transformações para variáveis categóricas que aparecem no estudo dos métodos de discriminação para misturas. Estas transformações são interessantes pois usam a distribuição multinomial para estudar individualmente cada uma das categorias geradas pelos distintos arranjos do vetor de variáveis categóricas. Para começar considere-se o caso de uma única variável.

> **Definição 2.1.4** Transformação de uma variável categórica em vetor binário. Dada x uma variável aleatória com c categorias, então x é substituida por (c - 1) variáveis binárias x_1, ... x_{c-1}, tais que se a j-ésima categoria de x for observada então, $x_j = 1$ e $x_i = 0\ \forall\ i \neq j$ i, j = 1, ..., c - 1. Quando a última categoria c for observada, todas as variáveis binárias serão iguais a zero.

Exemplo 2.1.1 Num estudo sobre a procedência dos alunos de pós-graduação da UNICAMP define-se a variável aleatória x = lugar de nascimento, com três categorias possíveis: São Paulo, outro estado do Brasil e fora do Brasil. Usando a transformação anterior x é substituida por duas variáveis binárias x_1 e x_2 tais que:

x	(x_1, x_2)
São Paulo	(1, 0)
outro estado	(0, 1)
fora do Brasil	(0, 0)

Depois da transformação, o vetor $(x_1, x_2)'$ representa o lugar de nascimento dos alunos de pós-graduação da UNICAMP. ■

Considere-se agora o caso de duas ou mais variáveis categóricas.

Definição 2.1.5 Transformação de variáveis categóricas em uma multinomial. Dados $x_1, x_2, \ldots, x_c$ variáveis aleatórias categóricas tais que a i-ésima variável x_i tem k_i categorias possíveis (i = 1, ..., c) então, o vetor $x = (x_1, \ldots, x_c)'$ que assume $k = \prod_{i=1}^{c} k_i$ valores distintos será substituido por outro vetor $(z_1, \ldots, z_k)'$ com distribuição multinomial de maneira tal que se a m-ésima categoria definida por x é observada, $z_m = 1$ e $z_j = 0 \ \forall \ j \neq m$; j, m = 1, 2, ..., k.

Exemplo 2.1.2 Sejam as variáveis categóricas x_1 = sexo com duas categorias e x_2 = faixa etária com três categorias: [20, 35[, [35, 50[e 50 ou mais anos.

Na notação da definição 2.1.5 tem-se que $k_1 = 2$ e $k_2 = 3$, logo, $k = (2)(3) = 6$ portanto, a multinomial $z = (z_1, \ldots, z_6)'$ é definida da seguinte maneira:

$(x_1, x_2)'$	$(z_1, z_2, \ldots, z_6)'$
(Masc, [20,35[)	(1, 0, 0, 0, 0, 0)
(Masc, [35,50[)	(0, 1, 0, 0, 0, 0)
(Masc, 50 ou mais)	(0, 0, 1, 0, 0, 0)
(Fem, [20, 35[)	(0, 0, 0, 1, 0, 0)
(Fem, [35, 50[)	(0, 0, 0, 0, 1, 0)
(Fem, 50 ou mais)	(0, 0, 0, 0, 0, 1)

Neste exemplo, cada combinação possível de sexo e faixa etária representada originalmente no vetor $(x_1, x_2)'$, está representada (depois da transformação), por algum valor do vetor $z' = (z_1, \ldots, z_6)$. Note-se que $z \sim \text{Multin}(1, p_1, \ldots, p_6)$ onde $p_1 = P(x_1 = \text{masc e } x_2 = [20,35[); \ldots, p_6 = P(x_1 = \text{fem e } x_2 = 50 \text{ ou mais})$ são tais que $\sum_{i=1}^{6} p_i = 1$. ■

A seguinte definição representa um caso particular da transformação acima (definição 2.1.5), no qual todas as variáveis categóricas são binárias. Sob esta hipótese as categorias multinomiais podem ser estabelecidas usando uma expressão algébrica apropriada.

Definição 2.1.6 Transformação de um vetor binário em multinomial. Dadas q variáveis aleatórias binárias $x_1, \ldots, x_q$, então, estas são expressadas como uma multinomial $z = (z_1, \ldots, z_{2^q})$ de maneira que cada arranjo do vetor binário $x' = (x_1, \ldots, x_q)$ determina uma categoria multinomial de forma única dada por $m = 1 + \sum_{i=1}^{q} x_1 2^{(i-1)}$ assim, se $(x_1, \ldots, x_q)$ for observado, então, $z_m = 1$ e $z_j = 0 \ \forall \ j \neq m$; $j, m = 1, \ldots, 2^q$.

A transformação anterior foi usada por Krzanowski (1975) quando

propôs um método para discriminar mistura de variáveis binárias e contínuas baseado no modelo de posição (ver cap.3 seção 3.2).

Exemplo 2.1.3 Sejam x_1, x_2, e x_3 variáveis binárias, usando a transformação da definição 2.1.6, estas variáveis são substituidas por uma multinomial de $2^3 = 8$ categorias.

$x = (x_1, x_2, x_3)$	m*	$(z_1, \ldots, z_8)$
$c_1 = (0, 0, 0)$	1	(1, 0, 0, 0, 0, 0, 0, 0)
$c_2 = (1, 0, 0)$	2	(0, 1, 0, 0, 0, 0, 0, 0)
$c_3 = (0, 1, 0)$	3	(0, 0, 1, 0, 0, 0, 0, 0)
$c_4 = (0, 0, 1)$	5	(0, 0, 0, 0, 1, 0, 0, 0)
$c_5 = (1, 1, 0)$	4	(0, 0, 0, 1, 0, 0, 0, 0)
$c_6 = (1, 0, 1)$	6	(0, 0, 0, 0, 0, 1, 0, 0)
$c_7 = (0, 1, 1)$	7	(0, 0, 0, 0, 0, 0, 1, 0)
$c_8 = (1, 1, 1)$	8	(0, 0, 0, 0, 0, 0, 0, 1)

(*) $m = 1 + \sum_{i=1}^{3} x_i 2^{(i-1)}$

neste caso $z = (z_1, \ldots, z_8) \sim$ Multin(1, $p_1, \ldots, p_8$) onde $p_1 = P(x=c_i)$ e $\sum_{i=1}^{8} p_i = 1$. ■

Em 1980, Krzanowski extendeu o método de discriminação proposto em 1975 para discriminar com mistura de variáveis categóricas e contínuas em geral. Esta extensão é simples usando o fato de que qualquer variável categórica pode ser transformada em um vetor binário conforme foi visto na definição 2.1.4.

O seguinte exemplo ilustrará a aplicação conjunta das definições 2.1.4 e 2.1.6 para transformar um vetor de variáveis

categóricas em vetor de binárias e depois em multinomial. Esta transformação que poderia ser feita diretamente usando a definição 2.1.5, é ilustrada aqui pois é assim que é usada em Krzanowski (1980) (ver subseção 3.2.3).

Exemplo 2.1.4 Sejam as variáveis x_1 = sexo e x_2 = faixa etária definidas no exemplo 2.1.2; x_1 tem 2 categorias, então x_1 é substituida por uma variável binária b_1 tal que se x_1 = feminino $\Rightarrow b_1 = 0$ e se x_1 = masculino $\Rightarrow b_1 = 1$; x_2 tem 3 categorias, então, x_2 é substituida por duas variáveis binárias b_2, b_3 tais que:

$$\begin{aligned} \text{se } x_2 = [20, 35[&\Rightarrow (b_2, b_3) = (1, 0) \\ x_2 = [35, 50[&\Rightarrow (b_2, b_3) = (0, 1) \\ x_2 = 50 \text{ ou mais} &\Rightarrow (b_2, b_3) = (0, 0) \end{aligned}$$

Portanto a informação que originalmente se obtinha do vetor categórico (x_1, x_2) agora está representada nos seis valores diferentes que assume o vetor binário $b = (b_1, b_2, b_3)$ da seguinte maneira:

(x_1, x_2)	(b_1, b_2, b_3)
(F, [20, 35[)	(0, 1, 0)
(F, [35, 50[)	(0, 0, 1)
(F, 50 ou mais)	(0, 0, 0)
(M, [20, 35[)	(1, 1, 0)
(M, [35, 50[)	(1, 0, 1)
(M, 50 ou mais)	(1, 0, 0)

Observe-se que os valores (1, 1, 1) e (0, 1, 1) do vetor binário não representam nenhuma combinação possível de sexo com faixa

etária, portanto, assume-se que P[b = (1, 1, 1)] = P[b = (0, 1, 1)] =0.

Agora, usando a definição 2.1.6 no vetor binário (b_1, b_2, b_3), estas variáveis são substituidas por uma multinomial com $2^3 = 8$ categorias de tal maneira que:

(x_1, x_2)	(b_1, b_2, b_3)	m*	$(z_1, \ldots, z_8)$
(F, [20, 35[)	(0, 1, 0)	3	(0, 0, 1, 0, 0, 0, 0, 0)
(F, [35, 50[)	(0, 0, 1)	5	(0, 0, 0, 0, 1, 0, 0, 0)
(F, 50 ou mais)	(0, 0, 0)	1	(1, 0, 0, 0, 0, 0, 0, 0)
(M, [20, 35[)	(1, 1, 0)	4	(0, 0, 0, 1, 0, 0, 0, 0)
(M, [35, 50[)	(1, 0, 1)	6	(0, 0, 0, 0, 0, 1, 0, 0)
(M, 50 ou mais)	(1, 0, 0)	2	(0, 1, 0, 0, 0, 0, 0, 0)

(*) $m = \sum_{i=1}^{3} b_i 2^{(i-1)} + 1$

O vetor $\mathbf{z} = (z_1, \ldots, z_8) \sim$ Multin(1, $p_1, \ldots, p_8$) onde,

$$p_7 = P(z_7 = 1) = P(b = (0, 1, 1)) = 0 \text{ e}$$
$$p_8 = P(z_8 = 1) = P(b = (1, 1, 1)) = 0,$$

pois, como foi visto antes, essas combinações das variáveis binárias não representam nenhuma combinação de sexo e faixa etária. ■

Para terminar esta seção será definida uma transformação muito similar à anterior (definição 2.1.6), porém diferente pois o vetor multinomial é formado por $r = 2^q - 1$ variáveis binárias sendo que uma delas foi omitida para evitar a dependência linear entre as componentes x_i's. Esta transformação foi usada por Vlachonikolis e Marriott (1982) para construir algumas modificações da Função Linear Discriminante de Fisher que melhoram seu desempenho em presença de misturas de variáveis.

Definição 2.1.7 q variáveis binárias $x_1, \ldots, x_q$ são substituidas por $r = 2^q - 1$ variáveis $z_1, \ldots, z_r$ tais que se $m = \sum_{i=1}^{q} x_i 2^{(i-1)} \in \{1, 2, \ldots, r\}$, então, $z_m = 1$ e $z_j = 0 \ \forall \ j \neq m$; $j = 1, \ldots, r$ e se $m = 0$ então $z_j = 0 \ \forall \ j = 1, \ldots, r$.

Observe-se que o vetor $(z_1, \ldots, z_r)'$ está formado por um subconjunto das variáveis do vetor multinomial z da definição 2.1.6, logo, pela propriedade 2.1.1 da distribuição multinomial, também é multinomial se a variável $z_0 = 1 - \sum_{j=1}^{r} z_j$ é considerada.

2.2 - DERIVAÇÃO DO VETOR DE MÉDIAS E MATRIZ DE VARIÂNCIA-COVARIÂNCIA.

Nesta seção é derivado o vetor de médias e a matriz de variância-covariância populacional e amostral de vetores mistos de três tipos:

i) Mistura de variáveis binárias e contínuas.

ii) Mistura de uma multinomial e variáveis contínuas.

iii) Mistura de uma multinomial, variáveis contínuas e as interações entre elas representadas pelos produtos cruzados das componentes multinomiais e as variáveis contínuas.

Na primeira parte será estudado o caso populacional. A notação usada será definida oportunamente e os resultados serão comentados e comparados entre os diferentes tipos de misturas. O caso amostral será tratado na segunda parte desta seção.

2.2.1 - Caso Populacional

a) Mistura de variáveis binárias e contínuas.

Para começar será considerado um vetor misto de variáveis binárias e contínuas:
Seja $\mathbf{w} = (\mathbf{x}', \mathbf{y}')'$ vetor aleatório misto onde:

$\mathbf{x} = (x_1, \ldots, x_q)'$ é um vetor aleatório binário

$\mathbf{y} = (y_1, \ldots, y_p)'$ é um vetor aleatório contínuo.

As componentes de $\mathbf{x}$ são binárias portanto este vetor assume 2^q valores diferentes, (conforme visto na definição 2.1.5), que serão chamados $c_1, c_2, \ldots, c_{2^q}$, assim, cada c_j representa algum dos possíveis arranjos do vetor binário.

Para facilitar o cálculo dos momentos considere as seguintes definições:

$S(x_i)$ = { conjunto de índices dos valores de $\mathbf{x}$ nos quais $x_i = 1$}
$= \{j \in \{1, \ldots, 2^q\} \mid c_j = (x_1, \ldots, x_{i-1}, 1, x_{i+1}, \ldots, x_q)\}$;
$i = 1, \ldots, q$

$S(x_i, x_j)$ = {conjunto de índices dos valores de $\mathbf{x}$ nos quais $x_i=1$ e $x_j=1$}
$= \{k \in \{1, \ldots, 2^q\} \mid c_k = (x_1, \ldots, x_{i-1}, 1, \ldots, x_{j-1}, 1, \ldots, x_q)\}$;
$i, j = 1, \ldots, q \; ; \; i \neq j.$

$p_j = P(\mathbf{x} = c_j)$ $j = 1, \ldots, 2^q$; probabilidade de que $\mathbf{x}$ assuma o valor c_j.

$\mu_j^{(k)} = E(y_j \mid \mathbf{x}=c_k)$, esperança condicional de y_j dado que $\mathbf{x}$ assume o

valor c_k , $k = 1, \ldots, 2^q$ $\quad j = 1, \ldots, p$

$\sigma_{ij}^{(k)} = \text{cov}\ (\ y_i,\ y_j | x=c_k)$, covariância condicional entre y_i e y_j dado que $\mathbf{x}$ assume o valor c_k, $k = 1, \ldots, 2^q$; $i, j = 1, \ldots, p$

Agora, a esperança de $\mathbf{w}$ está dada por:

$$E(w) = E \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \begin{pmatrix} E\mathbf{x} \\ E\mathbf{y} \end{pmatrix}$$

mas

$$E\mathbf{x} = E(\ x_1, \ldots, x_q)' = (Ex_1, \ldots, Ex_q)' \quad e$$

$$E\mathbf{y} = E\ (\ y_1, \ldots, y_p)' = (Ey_1, \ldots, Ey_p)'$$

$$E(\ x_i\ ) = 0\ P(\ x_i = 0\ ) + 1\ P(x_i = 1) = P(\ x_i = 1)$$

$$P(\ x_i = 1) = \sum_{j=1}^{2^q} P(\ x_i = 1 | \mathbf{x} = c_j)\ P(\ \mathbf{x} = c_j)$$

mas

$$P(\ x_i = 1 | \mathbf{x} = c_j) = \begin{cases} 1 \text{ se } j \in S(x_i) \\ 0 \text{ c.c.} \end{cases}$$

logo

$$P(\ x_i = 1) = \sum_{j \in S(x_i)} P(\ \mathbf{x} = c_j) = \sum_{j \in S(x_i)} p_j$$

$$\therefore E(x_i) = \sum_{j \in S(x_i)} p_j \qquad i = 1, \ldots, q \qquad (2.2.1)$$

Em (2.2.1) tem-se que a esperança de x_i é o somatório das probabilidades de que $\mathbf{x}$ assuma algum valor para o qual $x_i = 1$.

$$E(y_1) = \sum_{j=1}^{2^q} E(y_1 | x=c_j)\, P(x=c_j) = \sum_{j=1}^{2^q} \mu_1^{(j)} p_j$$

$$E(y_1) = \sum_{j=1}^{2^q} \mu_1^{(j)} p_j = \bar{\mu}_1 \qquad (2.2.2)$$

Note-se que a esperança de y_1, denotada por $\bar{\mu}_1$, representa uma média ponderada das esperanças condicionais de y_1 dados os valores de **x**, onde os pesos são as probabilidades de assumir esses valores respectivamente.

Portanto, de (2.2.1) e (2.2.2) tem-se que o vetor de médias do vetor aleatório misto **w** é dado por:

$$E(\mathbf{w}) = \left(\sum_{j \in S(x_1)} p_j, \ \ldots, \ \sum_{j \in S(x_q)} p_j, \ \bar{\mu}_1, \ \ldots, \ \bar{\mu}_p \right)$$

A seguir será calculada a matriz de variância-covariância de **w**:

Seja $\mathrm{cov}(\mathbf{w}) = \Omega = \begin{bmatrix} \Psi & \Delta \\ \Delta' & \Gamma \end{bmatrix}$, onde

$$\Psi = [\psi_{ij}] = [\mathrm{cov}(x_i, x_j)] \qquad i, j = 1, \ldots, q$$

$$\Delta = [\delta_{ij}] = [\mathrm{cov}(x_i, y_j)] \qquad i = 1, \ldots, q; \ j = 1, \ldots, p$$

$$\Gamma = [\gamma_{ij}] = [\mathrm{cov}(y_i, y_j)] \qquad i, j = 1, \ldots, p$$

Primeiramente será derivado o bloco Ψ correspondente às variáveis binárias:

$$\psi_{ij} = E(x_i x_j) - E(x_i)E(x_j) \qquad \text{se } i \neq j \qquad (2.2.3)$$

agora,

$$E(x_i x_j) = 1\ P(x_i=1,\ x_j=1) = \sum_{k=1}^{2^q} P(x_i=1,\ x_j=1 \mid x=c_k)p_k$$

mas

$$P(x_i=1,\ x_j=1 \mid x=c_k) = \begin{cases} 1 & \text{se } k \in S(x_i,\ x_j) \\ 0 & \text{c.c.} \end{cases}$$

logo:

$$E(x_i x_j) = \sum_{k \in S(x_i,x_j)} p_k \qquad i,j = 1, \ldots, q \qquad i \neq j \tag{2.2.4}$$

substituindo (2.2.1) e (2.2.4) em (2.2.3) tem-se:

$$\psi_{ij} = \sum_{k \in S(x_i,x_j)} p_k - \left(\sum_{k \in S(x_i)} p_k \right) \left(\sum_{k \in S(x_j)} p_k \right) \tag{2.2.5}$$

$i \neq j\ ;\ i, j = 1, \ldots, q$

Agora:

$$\psi_{ii} = \text{var}\ (x_i) = E(x_i^2) - [E(x_i)]^2 = P(x_i = 1) - [P(x_i = i)]^2$$

logo de (2.2.1),

$$\psi_{ii} = \left(\sum_{j \in S(x_i)} p_j \right) \left(1 - \sum_{j \in S(x_i)} p_j \right) \tag{2.2.6}$$

Derivando o bloco correspondente às covariâncias entre binárias e contínuas (matriz Δ), tem-se:

$$\delta_{ij} = E(x_i y_j) - E(x_i)E(y_j) \qquad i = 1, \ldots, q;\ j = 1, \ldots, p \tag{2.2.7}$$

$$E(x_i y_j) = \sum_{k=1}^{2^q} E(x_i y_j \mid x = c_k)p_k$$

mas

$$E(x_i y_j | \mathbf{x} = c_k) = \begin{cases} E(y_j | \mathbf{x} = c_k) = \mu_j^{(k)} & \text{se } k \in S(x_i) \\ 0 & \text{c.c.} \end{cases}$$

logo:

$$E(x_i y_j) = \sum_{k \in S(x_i)} \mu_j^{(k)} p_k \qquad i = 1, \ldots, q \; ; \quad j = 1, \ldots, p \tag{2.2.8}$$

de (2.2.8), (2.2.1) e (2.2.2) em (2.2.7) tem-se:

$$\delta_{ij} = \sum_{k \in S(x_i)} \mu_j^{(k)} p_k - \left(\sum_{k \in S(x_i)} p_k \right) \bar{\mu}_j$$

$$\delta_{ij} = \sum_{k \in S(x_i)} p_k \left(\mu_j^{(k)} - \bar{\mu}_j \right) \qquad \begin{matrix} i = 1, \ldots, q; \\ j = 1, \ldots, p \end{matrix} \tag{2.2.9}$$

em (2.2.9) observa-se que a $\text{cov}(x_i, y_j) = \delta_{ij}$ é uma média ponderada dos desvios entre as médias condicionais de y_j dado $\mathbf{x}$ e a média de y_j. Contribuem nestas médias unicamente os valores de $\mathbf{x}$ para os quais $x_i = 1$ e tem como coeficientes de ponderação as probabilidades de $\mathbf{x}$ assumir esses valores.

Para que a matriz cov(w) esteja totalmente definida falta conhecer as componentes de Γ, ou seja, a matriz de variância e covariância entre contínuas na presença de misturas, então:

$$\gamma_{ij} = E(y_i y_j) - E(y_i)\, E(y_j) \qquad i \neq j \qquad i, j = 1, \ldots, p$$

$$E(y_i y_j) = \sum_{k=1}^{2^q} E(y_i y_j | \mathbf{x} = c_k)\, p_k$$

mas

$$\text{cov}(y_i, y_j | \mathbf{x} = c_k) = \sigma_{ij}^{(k)} = E(y_i y_j | \mathbf{x} = c_k) - E(y_i | \mathbf{x} = c_k)\, E(y_j | \mathbf{x} = c_k)$$

(ver a definição de variância condicional em James, 1981 p.176)

como,

$$E(y_i y_j | x = c_k) = \sigma_{ij}^{(k)} + \mu_i^{(k)} \mu_j^{(k)} \qquad (2.2.10)$$

portanto

$$E(y_i y_j) = \sum_{k=1}^{2^q} \left(\sigma_{ij}^{(k)} + \mu_i^{(k)} \mu_j^{(k)} \right) p_k \qquad (2.2.11)$$

então

$$\gamma_{ij} = \sum_{k=1}^{2^q} \left(\sigma_{ij}^{(k)} + \mu_i^{(k)} \mu_j^{(k)} \right) p_k - \bar{\mu}_i \bar{\mu}_j$$

$$= \sum_{k=1}^{2^q} \sigma_{ij}^{(k)} p_k + \sum_{k=1}^{2^q} \left(\mu_i^{(k)} \mu_j^{(k)} - \bar{\mu}_i \bar{\mu}_j \right) p_k$$

$$\gamma_{ij} = \sum_{k=1}^{2^q} \sigma_{ij}^{(k)} p_k + \sum_{k=1}^{2^q} \left(\mu_i^{(k)} - \bar{\mu}_i \right) \left(\mu_j^{(k)} - \bar{\mu}_j \right) p_k$$

$$i,j = 1, \ldots, p \qquad (2.2.12)$$

em (2.2.12) a $\text{cov}(y_i, y_j) = \gamma_{ij}$, está escrita em função de duas parcelas que, identificados os valores que $\mathbf{x}$ assume como se fossem categorias, podem ser interpretadas como covariância dentro (primeira parcela) e a covariância entre as categorias (segunda parcela).

Se em (2.2.12) assume-se que $\sigma_{ij}^{(k)} = \sigma_{ij} \; \forall \; k = 1, \ldots, 2^q$, i.e., que as covariâncias condicionais de (y_i, y_j) são iguais para todos os valores de $\mathbf{x}$ então,

$$\gamma_{ij} = \sigma_{ij} + \sum_{k=1}^{2^q} \left(\mu_i^{(k)} - \bar{\mu}_i \right) \left(\mu_j^{(k)} - \bar{\mu}_j \right) p_k \qquad (2.2.13)$$

$$i,j = 1, \ldots, p;$$

Note-se que $var(y_i) = \gamma_{ii}$ é um caso particular de (2.2.12) assim,

$$\gamma_{ii} = \sum_{k=1}^{2^q} \sigma_{ii}^{(k)} p_k + \sum_{k=1}^{2^q} \left(\mu_i^{(k)} - \bar{\mu}_i \right)^2 p_k \qquad (2.2.14)$$

logo, se $\sigma_{ii}^{(k)} = \sigma_{ii} \quad \forall\ k = 1, \ldots, 2^q$ então,

$$\gamma_{ii} = \sigma_{ii} + \sum_{k=1}^{2^q} \left(\mu_i^{(k)} - \bar{\mu}_i \right)^2 p_k \quad i = 1, 2, \ldots, p \qquad (2.2.15)$$

Derivados portanto o vetor de médias e a matriz de variância-covariância da mistura de binárias com contínuas, ou seja, do vetor aqui denotado **w**, o próximo passo é estudar quais são as mudanças nas expressões encontradas sob as transformações de interesse definidas na seção 2.1.2.

b) Mistura de multinomial e contínuas.

Até agora trabalhou-se com $\mathbf{w} = (\mathbf{x}', \mathbf{y}')'$ porém, se a parte binária deste vetor é transformada usando a definição 2.1.7 tem-se um novo vetor misto:

$$\mathbf{w}^* = (\mathbf{z}', \mathbf{y}') \text{ onde } \mathbf{z} = (z_1, \ldots, z_r) \text{ e } r = 2^q - 1 \qquad (2.2.16)$$

Antes de começar a calcular os momentos de $\mathbf{w}^*$ é necessário definir a notação que será usada:

Seja z_0 a variável que completa o vetor multinomial **z**

$$z_0 = 1 - \sum_{j=1}^{r} z_j = \begin{cases} 1 & \text{se } z_j = 0 \ \forall\ j = 1, \ldots, r \\ 0 & \text{se } \exists\, z_j = 1 \quad j = 1, \ldots, r \end{cases} \qquad (2.2.17)$$

logo,

$$p_j = P(z_j = 1) \qquad j = 1, \ldots, r \qquad r = 2^q - 1 \qquad (2.2.18)$$

$$P(z_0 = 1) = 1 - P(z_0 = 0) = 1 - P(\exists\, z_j = 1, j=1,\ldots,r)$$
$$= 1 - \sum_{j=1}^{r} p_j = p_0 \qquad (2.2.19)$$

p_j é a probabilidade de observar a j-ésima categoria multinomial ($j = 0, 1, \ldots, r$).

$$\mu_j^{(m)} = E(y_j | z_m = 1) \quad m = 0, \ldots, r;\ j = 1, \ldots, p \qquad (2.2.20)$$

é a esperança condicional de y_j dado que foi observada a m-ésima categoria multinomial. Finalmente,

$$\sigma_{ij}^{(m)} = \text{cov}(y_i, y_j | z_m = 1) \quad m = 0, \ldots, r\ ;\ i,j = 1, \ldots, p \qquad (2.2.21)$$

representa a covariância condicional de y_i e y_j dado que foi observada a m-ésima categoria multinomial.

Logo, o vetor das médias populacionais da mistura $\mathbf{w}^*$ definida em (2.2.16) é, por definição:

$$E(\mathbf{w}^*) = (Ez_1, \ldots, Ez_r, Ey_1, \ldots, Ey_p)',$$

onde

$$E(z_m) = 0\ P(z_m = 0) + 1\ P(z_m = 1) = P(z_m = 1) = p_m \qquad m = 1, \ldots, r \qquad (2.2.22)$$

e

$$E(y_j) = \sum_{m=0}^{r} E(y_j | z_m = 1) P(z_m = 1) = \sum_{m=0}^{r} \mu_j p_m = \bar{\mu}_j \qquad (2.2.23)$$

Este resultado é equivalente ao obtido em (2.2.2). A diferença entre eles, é que em (2.2.23) se condiciona às categorias multinomiais enquanto que em (2.2.2) se condiciona aos valores que pode assumir o vetor **x**. Isto não faz diferença pois, após a transformação, cada valor de **x** está representado por uma única categoria multinomial.

Portanto, de (2.2.22) e (2.2.23) o vetor de médias buscado fica dado por:

$$E(\mathbf{w}^*) = (p_1, \ldots, p_r, \bar{\mu}_1, \ldots, \bar{\mu}_p)'$$

Para derivar a matriz de variância-covariância de $\mathbf{w}^*$, denotada por $\text{cov}(\mathbf{w}^*)$, considere a partição em blocos abaixo:

$$\text{cov}(\mathbf{w}^*) = \Omega^* = \begin{bmatrix} \Psi^* & \Delta^* \\ (\Delta^*)' & \Gamma^* \end{bmatrix} \qquad (2.2.24)$$

onde:

$$\Psi^* = [\psi^*_{ij}] = [\text{cov}(z_i, z_j)] \qquad j,i = 1, \ldots, r$$

$$\Delta^* = [\delta^*_{ij}] = [\text{cov}(z_i, y_j)] \qquad i = 1, \ldots, r;\ j = 1, \ldots, p$$

$$\Gamma^* = [\gamma^*_{ij}] = [\text{cov}(y_i, y_j)] \qquad i,j = 1, \ldots, p$$

Então, para derivar os elementos de Ψ^* tem-se:

$$\psi^*_{ij} = E(z_i z_j) - E(z_i)E(z_j) \quad i, j = 1, \ldots, r$$

mas

$$E(z_i z_j) = \begin{cases} 0 & \text{se } i \neq j \\ P(z_i = 1) & \text{se } i = j \end{cases}$$

então

$$\psi^{*}_{ij} = \begin{cases} -p_i p_j & \text{se } i \neq j \\ p_i(1 - p_i) & \text{se } i = j \end{cases} \quad i,j = 1, \ldots, r \qquad (2.2.25)$$

Este resultado era esperado pois ψ^{*}_{ij} é simplesmente a covariância entre duas componentes de um vetor multinomial.

Os elementos de Δ^{*} são obtidos a seguir, desde que:

$$\delta^{*}_{ij} = E(z_i y_j) - E(z_i)E(y_j),$$

e como,

$$E(z_i y_j) = \sum_{m=0}^{r} E(z_i y_j | z_m = 1)P(z_m = 1) = E(y_j | z_i = 1)p_i = \mu_j^{(i)} p_i \qquad (2.2.26)$$
$$i = 1, \ldots, r; \quad j = 1, \ldots, p$$

segue-se que

$$\delta^{*}_{ij} = \mu_j^{(i)} p_i - p_i \bar{\mu}_j = p_i(\mu_j^{(i)} - \bar{\mu}_j) \quad i = 1, \ldots, r; \; j = 1, \ldots, p \qquad (2.2.27)$$

ou seja, a covariância entre uma componente contínua y_j e uma categoria multinomial se expressa como o efeito na média da componente contínua, representado por uma diferença entre a média condicional de y_j na i-ésima categoria e a média de y_j, ponderada pela probabilidade de que a categoria i seja observada.

Observe-se que na expressão (2.2.9), que representa a covariância entre x_i e y_i, o resultado é parecido mas, nesse caso, tem-se um somatório de diferenças ponderadas para todos os valores de x nos quais $x_i = 1$.

As variâncias e covariâncias entre as componentes contínuas do vetor misto levando em conta a presença da multinomial, ou seja, γ^{*}_{ij}, são definidos como:

$$\gamma^{*}_{ij} = E(y_i y_j) - E(y_i)E(y_j), \qquad i, j = 1, \ldots, p$$

mas,

$$E(y_i y_j) = \sum_{m=0}^{r} E(y_i y_j | z_m = 1)p_m = \sum_{m=0}^{r}(\sigma^{(m)}_{ij} + \mu^{(m)}_i \mu^{(m)}_j)p_m \qquad (2.2.28)$$

logo:

$$\gamma^{*}_{ij} = \sum_{m=0}^{r}\sigma^{(m)}_{ij}p_m + \sum_{m=0}^{r} \mu^{(m)}_i \mu^{(m)}_j p_m - \bar{\mu}_i \bar{\mu}_j$$

que pode ser escrito como,

$$\gamma^{*}_{ij} = \sum_{m=0}^{r}\sigma^{(m)}_{ij}p_m + \sum_{m=0}^{r}(\mu^{(m)}_i - \bar{\mu}_i)(\mu^{(m)}_j - \bar{\mu}_j)p_m \qquad (2.2.29)$$
$$i,j = 1, \ldots, p$$

No caso particular em que se assume matrizes de covariância iguais em todas as categorias, (i.e. $\sigma^{(m)}_{ij} = \sigma_{ij} \quad \forall\ m = 0, \ldots, r$), tem-se:

$$\gamma^{*}_{ij} = \sigma_{ij} + \sum_{m=0}^{r}(\mu^{(m)}_i - \bar{\mu}_i)(\mu^{(m)}_j - \bar{\mu}_j)p_m \quad i, j = 1, \ldots, p \qquad (2.2.30)$$

As expressões (2.2.29) e (2.2.30) são totalmente equivalentes às obtidas para γ_{ij} em (2.2.12) e (2.2.13). Como já foi comentado antes, a diferença está em como são denotados as categorias geradas pelos distintos valores que assume o vetor binário **x**, quando se condicionam as esperanças. Novamente, é fácil identificar as parcelas destas expressões como revelando as componetes dentro e entre categorias.

c) Mistura de variáveis contínuas, multinomiais e interações entre elas.

Considerando agora o vetor misto espandido pelas componentes de produtos entre as variáveis multinomiais e contínuas, ou seja,

$$\mathbf{v} = (z_1, \ldots, z_r, y_1, \ldots, y_p, z_1y_1, \ldots, z_ry_p)' \qquad (2.2.31)$$

serão derivados o vetor de médias e a matriz de variância-covariância correspondentes. Este vetor é de interesse pois Vlachonikolis e Marriott (1982), construiram uma FLD modificada baseada já não nas variáveis originais $(x_1, \ldots, x_q, y_1, \ldots, y_p)$ senão no vetor $\mathbf{v}$ cujas componentes são, o vetor multinomial transformado $(z_1, \ldots, z_r)$, todas as variáveis contínuas $y_1, \ldots, y_p$, e os produtos da forma y_iz_j que representam as interações das variáveis contínuas com as distintas categorias multinomiais.

O vetor misto espandido, definido acima pode ser escrito como;

$$\mathbf{v} = (\mathbf{w}^{*'}, (\mathbf{z} \otimes \mathbf{y})')' \qquad (2.2.32)$$

onde $\mathbf{w}^* = (\mathbf{z}', \mathbf{y}')'$ já foi definido em (2.2.16) e $(\mathbf{z} \otimes \mathbf{y})$ representa o produto Kronecker dos vetores $\mathbf{z}$ e $\mathbf{y}$ (ver Mardia *et al.* 1979, p. 459).

O vetor das médias e matriz de variância-covariância populacionais de $\mathbf{v}$ ficam dados por:

$$E(\mathbf{v}) = [(E\mathbf{w}^*)', (E(\mathbf{z} \otimes \mathbf{y}))']' \qquad (2.2.33)$$

$$\mathrm{cov}(\mathbf{v}) = \begin{bmatrix} \Omega^* & \varepsilon \\ \varepsilon^T & \rho \end{bmatrix} \qquad (2.2.34)$$

os blocos são,

Ω^* = cov($\mathbf{w}^*$), já definida em (2.2.24) e

$$\varepsilon = [\mathrm{cov}(\mathbf{w}^*, \mathbf{z} \otimes \mathbf{y})] = \begin{bmatrix} \mathrm{cov}(\mathbf{z}, \mathbf{z} \otimes \mathbf{y}) \\ \hline \mathrm{cov}(\mathbf{y}, \mathbf{z} \otimes \mathbf{y}) \end{bmatrix} = \begin{bmatrix} \varepsilon_{11} \\ \hline \varepsilon_{21} \end{bmatrix}$$

onde:

$$\varepsilon_{11} = [(\varepsilon_{11})_{i,jk}] = [\mathrm{cov}(z_i, z_j y_k)] \qquad \begin{matrix} i,j = 1, \ldots, r \\ k = 1, \ldots, p \end{matrix}$$

$$\varepsilon_{21} = [(\varepsilon_{21})_{i,jk}] = [\mathrm{cov}(y_i, z_j y_k)] \qquad \begin{matrix} i,k = 1, \ldots, p \\ j = 1, \ldots, r \end{matrix}$$

$$\rho = \mathrm{cov}(\mathbf{z} \otimes \mathbf{y}) = [\rho_{ij,kl}] = [\mathrm{cov}(z_i y_j, z_k y_l)] \qquad \begin{matrix} i,k = 1, \ldots, r \\ j,l = 1, \ldots, p \end{matrix}$$

Para calcular o vetor de médias e a matriz de variância - covariância do vetor $\mathbf{v}$ não é necessário definir uma nova notação, continuará sendo usada a notação definida para calcular os momentos de $\mathbf{w}^*$.

Em (2.2.33) tem-se que:

$$E\mathbf{v} = ((E\mathbf{w}^*)', (E\ \mathbf{z} \otimes \mathbf{y})')' = ((E\mathbf{z})', (E\mathbf{y})', (E\ \mathbf{z} \otimes \mathbf{y})')'$$

ou seja,

$$E\mathbf{v} = (Ez_1, \ldots, Ez_r, Ey_1, \ldots, Ey_p, Ez_1y_1, \ldots, Ez_ry_p)' \qquad (2.2.35)$$

Observe-se que todas as componentes de Ev podem ser recuperadas dos cálculos feitos anteriormente, assim:

de (2.2.22) $E(z_i) = p_i$ $\qquad i = 1, \ldots, r$

de (2.2.23) $E(y_i) = \sum_{m=0}^{r} \mu_i^{(m)} p_m = \bar{\mu}_i \qquad i = 1, \ldots, p$

de (2.2.26) $E(z_i y_j) = \mu_j^{(i)} p_i \qquad i = 1, \ldots, r \quad j = 1, \ldots, p$

portanto, Ev está totalmente definida.

Note-se ainda, que em (2.2.34), a parcela Ω^* da cov(v) também é conhecida pois suas componentes foram definidas nas expressões (2.2.25) à (2.2.30). Trata-se agora de derivar os blocos da matriz da variância covariância que envolvem a parte expandida, ou seja, ε e ρ. As componentes de ε estão dadas por:

$$(\varepsilon_{11})_{1,jk} = E(z_i z_j y_k) - E(z_i)E(z_j y_k)$$

mas

$$E(z_i z_j y_k) = \sum_{m=0}^{r} E(z_i z_j y_k | z_m = 1) p_m = \begin{cases} 0 & \text{se } i \neq j \\ E(y_k | z_i = 1) p_i = \mu_k^{(i)} p_i & \text{se } i = j \end{cases} \tag{2.2.36}$$

então:

$$(\varepsilon_{11})_{i,jk} = \begin{cases} -p_i p_j \mu_k^{(j)} & \text{se } i \neq j \\ \mu_k^{(j)} p_j (1 - p_j) & \text{se } i = j \end{cases} \quad \begin{matrix} i, j = 1, \ldots, r \\ k = 1, \ldots, p \end{matrix} \tag{2.2.37}$$

Agora:

$$(\varepsilon_{21})_{i,jk} = E(y_i z_j y_k) - E(y_i)E(z_j y_k) \tag{2.2.38}$$

$$E(y_i z_j y_k) = \sum_{m=0}^{r} E(y_i z_j y_k / z_m = 1) p_m = E(y_i y_k / z_j = 1) p_j$$

analogamente a (2.2.10) tem-se

$$E(y_i y_k | z_j = 1) = (\sigma_{ik}^{(j)} + \mu_i^{(j)}\mu_k^{(j)})p_j \qquad i, k = 1, \ldots, p; \quad j = 1, \ldots, r \tag{2.2.39}$$

Portanto

$$(\varepsilon_{21})_{i,jk} = (\sigma_{ik}^{(j)} + \mu_i^{(j)}\mu_k^{(j)})p_j - \bar{\mu}_i\mu_k^{(j)}p_j$$

ou ainda,

$$(\varepsilon_{21})_{i,jk} = [\sigma_{ik}^{(j)} + (\mu_i^{(j)} - \bar{\mu}_i)\mu_k^{(j)}]p_j \qquad i, k = 1, \ldots, p; \quad j = 1, \ldots, r \tag{2.2.40}$$

Para expressar as componentes da matriz $\rho = \text{cov}(z \otimes y)$ escolhe-se um elemento e escreve-se como,

$$\rho_{ij,kl} = E(z_i y_j z_k y_l) - E(z_i y_j)E(z_k y_l) \tag{2.2.41}$$

mas,

$$E(z_i y_j z_k y_l) = \sum_{m=0}^{r} E(z_i y_j z_k y_l | z_m = 1)p_m \tag{2.2.42}$$

com,

$$E(z_i y_j z_k y_l | z_m = 1) = \begin{cases} 0 & \text{se } i \neq k \\ E(z_i^2 y_j y_l | z_m = 1) & \text{se } i = k \end{cases}$$

isto pois z_i e z_k representam categorias multinomiais, logo, se $i \neq k$ então $z_l = 0$ ou $z_k = 0$.

Portanto (2.2.42) se reduz a,

$$E(z_i y_j z_k y_l) = \begin{cases} 0 & \text{se } i \neq k \\ \sum_{m=0}^{r} E(z_i^2 y_j y_l | z_m = 1)p_m = E(y_j y_l | z_i = 1)p_i & \text{se } i = k \end{cases}$$

e de (2.2.39) segue-se que,

$$E(z_i y_j,\ z_k y_l) = \begin{cases} 0 & \text{se } i \neq k \\ (\sigma_{jl}^{(i)} + \mu_j^{(i)}\mu_l^{(i)})p_i & \text{se } i = k \end{cases} \qquad i,\ k = 1,\ \ldots,\ r \tag{2.2.43}$$

logo, de (2.2.26) e (2.2.43) em (2.2.41),

$$\rho_{ij,kl} = \begin{cases} -\ p_i p_k \mu_j^{(i)} \mu_l^{(i)} & \text{se } i \neq k \\ (\sigma_{jl}^{(i)} + \mu_j^{(i)}\mu_l^{(i)})p_i - p_i^2\ \mu_j^{(i)}\mu_l^{(i)} & \\ \quad = \sigma_{jl}^{(i)} p_i + (1 - p_i)\ p_i \mu_j^{(i)} \mu_l^{(i)} & \text{se } i = k \end{cases} \tag{2.2.44}$$

a expressão (2.2.44) vale para i, k = 1, ..., r e j, l = 1, ..., p

Caracterizadas populacionalmente as médias e matrizes de variância-covariância populacionais dos vetores mistos $\mathbf{w}$, $\mathbf{w}^*$ e $\mathbf{v}$, o passo seguinte é estudar os estimadores destes momentos, obtidos a partir de uma amostra aleatória de tamanho n.

2.2.2 - Caso Amostral

Nesta subseção os estimadores usuais de momentos do vetor de médias e da matriz de variância-covariância, que para o caso contínuo podem ser encontrados por exemplo em Mardia *et al.*, 1975 p. 10, são extendidos para o caso de observações sobre um vetor aleatório misto. Os momentos amostrais são derivados para cada caso tratado na subseção anterior, de forma vetorial ou matricial, como operações sobre a matriz de dados para tornar futuro desenvolvimento de programas mais fácil, facilitar interpretações geométricas e primordialmente explicitar a natureza das variâncias e covariâncias sob mistura.

a) Mistura de variáveis binárias e contínuas.

Em primeiro lugar será tratado o vetor misto $\mathbf{w} = (x_1, \ldots, x_q, y_1, \ldots, y_p)'$, definido na subseção 2.2.1 parte (a).

Neste caso a matriz de dados que representa uma amostra aleatória de tamanho n de observações de **w**, ou simplemente as observações escreve-se como

$$\mathbf{W}_{n \times (p+q)} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1q} & y_{11} & \cdots & y_{1p} \\ x_{21} & x_{22} & \cdots & x_{2q} & y_{21} & \cdots & y_{2p} \\ \vdots & \vdots & & \vdots & \vdots & & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nq} & y_{n1} & \cdots & y_{np} \end{bmatrix} \qquad (2.2.45)$$

o estimador da média de **w** expresso em função da matriz de dados acima é,

$$\hat{E(\mathbf{w})} = \bar{\mathbf{w}} = \frac{1}{n} \mathbf{W}'\mathbf{1}_n \qquad (2.2.46)$$

onde

$$\mathbf{1}'_n = (1,1, \ldots, 1)_{1 \times n} \qquad (2.2.47)$$

assim,

$$\bar{\mathbf{w}} = \frac{1}{n} \begin{bmatrix} x_{11} & x_{21} & \cdots & x_{n1} \\ \vdots & & & \vdots \\ x_{1q} & x_{2q} & \cdots & x_{nq} \\ y_{11} & y_{21} & \cdots & y_{n1} \\ \vdots & & & \vdots \\ y_{1p} & y_{2p} & \cdots & y_{np} \end{bmatrix} \begin{bmatrix} 1 \\ 1 \\ \cdot \\ \cdot \\ \cdot \\ \\ 1 \end{bmatrix}$$

$$\bar{\mathbf{w}} = \frac{1}{n} \left(\sum_{i=1}^{n} x_{i1}, \ldots, \sum_{i=1}^{n} x_{iq}, \sum_{j=1}^{n} y_{i1}, \ldots, \sum_{i=1}^{n} y_{ip} \right)' \qquad (2.2.48)$$

mas

$$\sum_{i=1}^{n} x_{ij} = m_j \; ; \quad j = 1, \ldots, q \qquad (2.2.49)$$

onde m_j representa o número de observações da amostra$_q$ de tamanho n, nas quais $x_j = 1$ (j = 1, ..., q). (Observe-se que $\sum_{j=1} m_j \neq n$, pois trata-se de q binárias não caracterizando portanto categorias mutuamente exclusivas). Por outro lado, $\sum_{i=1}^{n} y_{ik} = \bar{n}y_k$, isto é, n vezes a média amostral da variável y_k (k = 1, ..., p).

Ou seja, o vetor de médias da mistura independe para cada tipo de variável da presença do outro tipo, e fica dado por,

$$\bar{w} = \left(\frac{m_1}{n}, \frac{m_2}{n}, \ldots, \frac{m_q}{n}, \bar{y}_1, \ldots, \bar{y}_p \right)' \qquad (2.2.50)$$

O estimador da matriz de variância-covariância de **w** é S_w que expresso em função da matriz de dados **W** fica,

$$\hat{\Omega} = S_w^{(1)} = \frac{1}{n} \mathbf{W}' \mathbf{H} \mathbf{W} \quad \text{com } \mathbf{H} = I_n - \frac{1}{n} 1_n 1_n' \qquad (2.2.51)$$

logo,

$$S_w = \frac{1}{n} \begin{bmatrix} x_{11} & x_{12} & \ldots & x_{n1} \\ \vdots & & & \vdots \\ x_{1q} & x_{2q} & \ldots & x_{nq} \\ y_{11} & y_{21} & \ldots & y_{n1} \\ \vdots & & & \vdots \\ y_{1p} & y_{2p} & \ldots & y_{np} \end{bmatrix} \begin{bmatrix} 1-1/n & -1/n & \ldots & -1/n \\ 1/n & 1-1/n & \ldots & -1/n \\ . & . & & . \\ . & . & & . \\ . & . & & . \\ -1/n & -1/n & \ldots & 1-1/n \end{bmatrix} \begin{bmatrix} x_{11} & x_{12} & \ldots & y_{11} & \ldots & y_{1p} \\ x_{21} & x_{22} & \ldots & y_{21} & \ldots & y_{2p} \\ . & . & & . & & . \\ . & . & & . & & . \\ . & . & & . & & . \\ x_{n1} & x_{n2} & \ldots & y_{n1} & \ldots & y_{np} \end{bmatrix}$$

que depois de multiplicar as matrizes pode ser escrito como,

---

(1) S_w representa o estimador de máxima verossimilhança. O estimador não viciado da matriz de variância-covariância é dada por 1/(n-1)W'HW.

$$nS_w = \left[\begin{array}{ccc|ccc} \sum_{i=1}^{n} x_{i1}^2 - \frac{m_1^2}{n} & \cdots & \sum_{i=1}^{n} x_{i1}x_{iq} - \frac{m_1 m_q}{n} & \sum_{i=1}^{n} x_{i1}y_{i1} - m_1\bar{y}_1 & \cdots & \sum_{i=1}^{n} x_{i1}y_{ip} - m_1\bar{y}_p \\ & \ddots & \vdots & & \ddots & \vdots \\ & & \sum_{i=1}^{n} x_{iq}^2 - \frac{m_q^2}{n} & & & \sum_{i=1}^{n} x_{iq}y_{ip} - m_q\bar{y}_p \\ \hline & & & \sum_{i=1}^{n} y_{i1}^2 - n\bar{y}_1^2 & \cdots & \sum_{i=1}^{n} y_{i1}y_{ip} - n\bar{y}_1\bar{y}_p \\ & & & & \ddots & \vdots \\ & & & & & \sum_{i=1}^{n} y_{ip}^2 - n\bar{y}_p^2 \end{array}\right] \qquad (2.2.52)$$

Para simplificar a expressão anterior recordemos que $x_{ij} = \{1, 0\}$, portanto

$$\sum_{i=1}^{n} x_{ij}^2 = \sum_{i=1}^{n} x_{ij} = m_j \; ; \; j = 1, \ldots, q \qquad (2.2.53)$$

(ver 2.2.49). Agora

$$x_{ij}x_{ik} = \begin{cases} 1 & \text{se } x_{ij} = 1 \text{ e } x_{ik} = 1 \\ 0 & \text{c.c.} \end{cases} \qquad (2.2.54)$$

então definindo,

m_{kj} = número de observações na amostra de tamanho n, nas quais $x_j = 1$ e $x_k = 1$ $(j \neq k;\ j,k = 1, \ldots, q)$

de (2.2.54) tem-se que

$$\sum_{i=1}^{n} x_{ij}x_{ik} = m_{jk} \qquad j \neq k \; ; \; j,k = 1, \ldots, q \qquad (2.2.55)$$

Substituindo, (2.2.53) e (2.2.55) em (2.2.52); chega-se a seguinte

expressão para a matriz de dispersão amostral de **w**, $\mathbf{S}_w$;

$$\mathbf{S}_w = \left[\begin{array}{ccc|ccc} \frac{m_1}{n}\left(1-\frac{m_1}{n}\right) & \cdots & -\frac{m_{1q}}{n}-\frac{m_1 m_q}{n} & \frac{\sum_{i=1}^{n} x_{11} y_{11} - m_1 \bar{y}_1}{n} & \cdots & \frac{\sum_{i=1}^{n} x_{11} y_{1p} - m_1 \bar{y}_p}{n} \\ & \ddots & \vdots & & \ddots & \vdots \\ & & \frac{m_q}{n}\left(1-\frac{m_q}{n}\right) & & & \frac{\sum_{i=1}^{n} x_{iq} y_{ip} - m_q \bar{y}_p}{n} \\ \hline & & & s_{y_1 y_1} & s_{y_1 y_2} \cdots & s_{y_1 y_p} \\ & & & & s_{y_2 y_2} \cdots & s_{y_2 y_p} \\ & & & & \ddots & \vdots \\ & & & & & s_{y_p y_p} \end{array}\right]$$

onde

$$s_{y_i y_j} = \hat{\gamma}_{ij} = \frac{\sum_{k=1}^{n} y_{ki} y_{kj} - n\bar{y}_i \bar{y}_j}{n} \qquad i,j = 1, \ldots, p \qquad (2.2.56)$$

é a covariância amostral entre as variáveis y_i e y_j.

Note que a matriz de variância e covariância na presença de mistura de variáveis ganha novo significado, neste caso os blocos correspondentes às interrelações dentro de cada tipo de componente permanecem inalterados, mas há um novo bloco interligando e levando em conta a presença simultânea dos dois tipos de variáveis sobre os mesmos indivíduos. As técnicas multivariadas discretas ou contínuas passam então a ganhar especificidade pela nova natureza das interrelações.

A seguir serão calculados os estimadores da média e matriz de dispersão do vetor misto $\mathbf{v} = (\mathbf{z}', \mathbf{y}', (\mathbf{z} \otimes \mathbf{y})')'$, definido em

(2.2.31) e imediatamente depois, serão obtidos os estimadores correspondentes para o vetor $\mathbf{w}^* = (\mathbf{z}', \mathbf{y}')'$. A razão pela qual a ordem do tratamento dos vetores $\mathbf{w}^*$ e $\mathbf{v}$ é invertida em relação à seção anterior, é aproveitar o fato de que $\mathbf{w}^*$ pode ser obtido truncando a componente $(z \otimes y)$ do vetor $\mathbf{v}$ (i.e., os elementos da forma $z_i y_j$) portanto, para obter a matriz de dados correspondente a $\mathbf{w}^*$ basta suprimir da matriz de dados de $\mathbf{v}$ as colunas que representam as observações das variáveis $z_i y_j$ ($i = 1, \ldots, r$; $j = 1, \ldots, p$). Por outro lado, o vetor de médias e a matriz de variância-covariância de $\mathbf{w}^*$ são parcelas do vetor de médias e matriz de variância-covariância de $\mathbf{v}$, conforme foi visto em (2.2.33) e (2.2.34), logo é fácil obtê-los a partir dos estimadores para $\mathbf{v}$.

b) Mistura de variáveis binárias, contínuas e as interações entre elas.

Seja a matriz de dados:

$$\mathbf{V} = \left[\begin{array}{ccc|ccc|ccc} z_{11} & z_{12} & z_{1r} & y_{11} & y_{1p} & (z_1y_1)_1 & & (z_ry_p)_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \cdots & \vdots \\ z_{n1} & z_{n2} & z_{nr} & y_{n1} & y_{np} & (z_1y_1)_n & & (z_ry_p)_n \end{array}\right]_{n \times s} \qquad (2.2.57)$$

onde $s = r + p + pr = 2^q - 1 + p + p(2^q - 1) = 2^q(p + 1) - 1$

Agora supondo que, das n observações que compõem a matriz $\mathbf{V}$, n_i ($i = 0, 1, \ldots, r$) pertencem à i-ésima categoria multinomial, então, tem-se que: $\sum_{j=0}^{r} n_j = n$.

Note-se que é possível reordenar as linhas da matriz $\mathbf{V}$ e colocá-las de maneira tal que todas as observações de uma mesma categoria fiquem juntas formando blocos. Assim, $\mathbf{V}$ pode ser escrita da seguinte maneira:

$$\mathbf{V} = [\mathbf{V}^{(1)'} | \mathbf{V}^{(2)'} | \ \ldots \ | \mathbf{V}^{(r)'} | \mathbf{V}^{(0)'}]^{T} \qquad (2.2.58)$$

onde $\mathbf{V}^{(j)}_{(njxs)}$ representa o bloco de todas as obsevações da amostra que pertencem a j-ésima categoria multinomial (j = 0, 1, ..., r).

O objetivo de reordenar as linhas de **V** é aproveitar o fato de que, dentro de cada bloco, as observações têm uma forma particular que simplifica muito o cálculo dos estimadores. Por exemplo, para um vetor $\tilde{v}= (z_1, z_2, z_3, y_1, y_2, z_1y_1, z_1y_2, z_2y_1, z_2y_2, z_3y_1, z_3y_2)'$ a matriz de dados reordenada seria como segue:

$$\tilde{\mathbf{V}} = \begin{array}{c} \\ n_1 \\ \\ \\ \\ n_2 \\ \\ \\ \\ n_3 \\ \\ \\ \\ n_0 \\ \\ \\ \end{array} \begin{array}{c} \begin{array}{cccccccccccc} z_1 & z_2 & z_3 & y_1 & y_2 & z_1y_1 & z_1y_2 & z_2y_1 & z_2y_2 & z_3y_1 & z_3y_2 \end{array} \\ \left[\begin{array}{cccccccccccc} 1 & 0 & 0 & y_{11}^{(1)} & y_{12}^{(1)} & y_{11}^{(1)} & y_{12}^{(1)} & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & 0 & 0 & y_{n_1 1}^{(1)} & y_{n_1 2}^{(1)} & y_{n_1 1}^{(1)} & y_{n_1 2}^{(1)} & 0 & 0 & 0 & 0 \\ \hline 0 & 1 & 0 & y_{11}^{(2)} & y_{12}^{(2)} & 0 & 0 & y_{11}^{(2)} & y_{12}^{(2)} & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 1 & 0 & y_{n_2 1}^{(2)} & y_{n_2 2}^{(2)} & 0 & 0 & y_{n_2 1}^{(2)} & y_{n_2 2}^{(2)} & 0 & 0 \\ \hline 0 & 0 & 1 & y_{11}^{(3)} & y_{12}^{(3)} & 0 & 0 & 0 & 0 & y_{11}^{(3)} & y_{12}^{(3)} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 1 & y_{n_3 1}^{(3)} & y_{n_3 2}^{(3)} & 0 & 0 & 0 & 0 & y_{n_3 1}^{(3)} & y_{n_3 2}^{(3)} \\ \hline 0 & 0 & 0 & y_{11}^{(0)} & y_{12}^{(0)} & 0 & 0 & 0 & 0 & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & y_{n_0 1}^{(0)} & y_{n_0 2}^{(0)} & 0 & 0 & 0 & 0 & 0 & 0 \end{array}\right] \end{array} \qquad (2.2.59)$$

onde $y_{ij}^{(k)}$ representa a i-ésima observação da variável y_j na categoria multinomial k (i = 1, ..., n_k ; j = 1, 2 ; k = 0, 1, 2, 3).

A matriz de dados anterior pode ser escrita de uma maneira mais simples se a seguinte notação é usada:

$\mathbf{1}_{n_j}$ = (1, 1, ..., 1)' vetor (n_j x 1) que foi definido em (2.2.47),

$\mathbf{O}_{n_j \times p}$: matriz nula de dimensão (n_j x p) (2.2.60)

$\mathbf{O}_{n_j}$: vetor nulo (n_j x 1) (2.2.61)

$$\mathbf{Y}^{(j)} = \begin{bmatrix} y_{11}^{(j)} & y_{12}^{(j)} & \cdots & y_{1p}^{(j)} \\ \vdots & \vdots & & \vdots \\ y_{n_j 1}^{(j)} & y_{n_j 2}^{(j)} & & y_{n_j p}^{(j)} \end{bmatrix}_{n_j \times p} \qquad (2.2.62)$$

$\mathbf{Y}^{(j)}$ matriz de observações do vetor $\mathbf{y}$ na j-ésima categoria multinomial.

Com esta notação, $\tilde{\mathbf{V}}$ em (2.2.59) é:

$$\tilde{\mathbf{V}} = \begin{array}{c} \begin{array}{ccccccc} z_1 & z_2 & z_3 & y & z_1 y & z_2 y & z_3 y \end{array} \\ \left[\begin{array}{ccc:c:c:c:c} \mathbf{1}_{n_1} & \mathbf{O}_{n_1} & \mathbf{O}_{n_1} & \mathbf{Y}^{(1)} & \mathbf{Y}^{(1)} & \mathbf{O}_{n_1 \times 2} & \mathbf{O}_{n_1 \times 2} \\ \hdashline \mathbf{O}_{n_2} & \mathbf{1}_{n_2} & \mathbf{O}_{n_2} & \mathbf{Y}^{(2)} & \mathbf{O}_{n_2 \times 2} & \mathbf{Y}^{(2)} & \mathbf{O}_{n_2 \times 2} \\ \hdashline \mathbf{O}_{n_3} & \mathbf{O}_{n_3} & \mathbf{1}_{n_3} & \mathbf{Y}^{(3)} & \mathbf{O}_{n_3 \times 2} & \mathbf{O}_{n_3 \times 2} & \mathbf{Y}^{(3)} \\ \hdashline \mathbf{O}_{n_0} & \mathbf{O}_{n_0} & \mathbf{O}_{n_0} & \mathbf{Y}^{(0)} & \mathbf{O}_{n_0 \times 2} & \mathbf{O}_{n_0 \times 2} & \mathbf{O}_{n_0 \times 2} \end{array}\right] \end{array}$$

Generalizando o resultado anterior para o vetor $v = (z_1, \ldots, z_r, y_1, \ldots, y_p, z_1y_1, z_1y_2, \ldots, z_ry_p)'$ e usando a notação acima definida, tem-se que a matriz de dados V, em (2.2.57), pode ser escrita como segue:

$$V = \begin{array}{c} \begin{array}{cccccccc} z_1 & z_2 & & z_r & y & z_1y & z_2y & z_ry \end{array} \\ \left[\begin{array}{cccc|c|c|c|c} 1_{n_1} & 0_{n_1} & \ldots & 0_{n_r} & Y^{(1)}_{n_1 x\, p} & Y^{(1)}_{n_1 x\, p} & 0_{n_1 x\, p} & \ldots & 0_{n_1 x\, p} \\ \hline 0_{n_2} & 1_{n_2} & \ldots & 0_{n_2} & Y^{(2)}_{n_2 x\, p} & 0_{n_2 x\, p} & Y^{(2)}_{n_2 x\, p} & \ldots & 0_{n_2 x\, p} \\ \hline \vdots & \vdots & & \vdots & \vdots & \vdots & \vdots & & \vdots \\ \hline 0_{n_r} & 0_{n_r} & \ldots & 1_{n_r} & Y^{(r)}_{n_r x\, p} & 0_{n_r x\, p} & 0_{n_r x\, p} & \ldots & Y^{(r)}_{n_r x\, p} \\ \hline 0_{n_0} & 0_{n_0} & \ldots & 0_{n_0} & Y^{(0)}_{n_0 x\, p} & 0_{n_0 x\, p} & 0_{n_0 x\, p} & \ldots & 0_{n_0 x\, p} \end{array}\right]_{n \times s} \end{array}$$

(2.2.63)

O vetor de médias amostrais de v, como função da matriz de dados V é dado por:

$$\bar{v} = \frac{1}{n} V'1_n = \frac{1}{n}(1_n' V)'$$

ou seja,

$$\bar{v} = \frac{1}{n}[n_1, n_2, \ldots, n_r \mid \sum_{m=0}^{r} 1_{n_m}' Y^{(m)} \mid 1_{n_1}' Y^{(1)} \mid 1_{n_2}' Y^{(2)} \mid \ldots \mid 1_{n_r}' Y^{(r)}]'$$

(2.2.64)

mas,

$$1'_{n_m} Y^{(m)} = (1, \ldots, 1) \begin{bmatrix} y_{11}^{(m)} \cdots y_{p1}^{(m)} \\ \vdots \quad \vdots \\ y_{n_m 1}^{(m)} \cdots y_{n_m p}^{(m)} \end{bmatrix}$$

$$= n_m (\bar{y}_1^{(m)}, \ldots, \bar{y}_p^{(m)}) = n_m \bar{y}^{(m)'} \qquad (2.2.65)$$

onde $\bar{y}^{(m)}$ é a média amostral do vetor **y** na categoria m, baseada nas n_m observações nessa categoria (m = 0, 1, ..., r) logo a média amostral do vetor **y** é:

$$\bar{y} = \frac{1}{n} \sum_{m=0}^{r} n_m \bar{y}^{(m)} \qquad (2.2.66)$$

De (2.2.64) a (2.2.66) segue a expressão do vetor de médias de **v** como;

$$\bar{v} = \Big[\underbrace{\frac{n_1}{n}, \ldots, \frac{n_r}{n}}_{1 \times r} \Big| \underbrace{\bar{y}'}_{1 \times p} \Big| \underbrace{\frac{n_1 \bar{y}^{(1)'}}{n}}_{1 \times p} \Big| \underbrace{\frac{n_2 \bar{y}^{(2)'}}{n}}_{1 \times p} \Big| \cdots \Big| \underbrace{\frac{n_r \bar{y}^{(r)'}}{n}}_{1 \times p} \Big]'_{1 \times s}$$

(2.2.67)

Para derivar a matriz de variância-covariância amostral de **v** através da matriz de dados como construida em (2.2.63), basta explicitar,

$$S_v = \frac{1}{n} V' \, H \, V = \frac{1}{n} V'(I - \frac{1}{n} 1_n 1'_n) V \qquad (2.2.68)$$

Primeiramente é necessário paticionar a matriz de projeção $H = (I - 1/n 1_n 1'_n)$ de maneira compatível com a forma de blocos de **V** para que o produto das duas seja possível, assim;

$$
\mathbf{H}_{n\times n}=\left[\begin{array}{c|c|c|c}
\begin{matrix} 1-1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & 1-1/n \end{matrix} & \begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} & \cdots & \begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} \\ \hline
\begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} & \begin{matrix} 1-1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & 1-1/n \end{matrix} & \cdots & \begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} \\ \hline
\vdots & \vdots & \ddots & \vdots \\ \hline
\begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} & \begin{matrix} -1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & -1/n \end{matrix} & \cdots & \begin{matrix} 1-1/n & \cdots & -1/n \\ \vdots & \ddots & \vdots \\ -1/n & \cdots & 1-1/n \end{matrix}
\end{array}\right]
\begin{array}{l} \} n_1 \\ \} n_2 \\ \\ \\ \} n_0 \end{array}
$$

$$
\underbrace{\qquad}_{n_1} \quad \underbrace{\qquad}_{n_2} \quad \cdots \quad \underbrace{\qquad}_{n_0}
$$

recuperando a notação compacta e escrevendo os blocos em forma mais simples,

$$
\mathbf{H} = \left[\begin{array}{c|c|c|c}
\mathbf{H}_{n_1} & -\frac{1}{n}\,1_{n_1}1'_{n_2} & \cdots & -\frac{1}{n}\,1_{n_1}1'_{n_r} & -\frac{1}{n}\,1_{n_1}1'_{n_0} \\ \hline
-\frac{1}{n}\,1_{n_2}1'_{n_1} & \mathbf{H}_{n_2} & \cdots & -\frac{1}{n}\,1_{n_2}1'_{n_r} & -\frac{1}{n}\,1_{n_2}1'_{n_0} \\ \hline
\vdots & \vdots & \ddots & \vdots & \vdots \\ \hline
-\frac{1}{n}\,1_{n_r}1'_{n_1} & -\frac{1}{n}\,1_{n_r}1'_{n_2} & \cdots & \mathbf{H}_{n_r} & -\frac{1}{n}\,1_{n_r}1'_{n_0} \\ \hline
-\frac{1}{n}\,1_{n_0}1'_{n_1} & -\frac{1}{n}\,1_{n_0}1'_{n_2} & \cdots & -\frac{1}{n}\,1_{n_0}1'_{n_r} & \mathbf{H}_{n_0}
\end{array}\right]
\begin{array}{l} \} n_1 \\ \} n_2 \\ \\ \} n_r \\ \} n_0 \end{array}
\qquad (2.2.69)
$$

onde $\mathbf{H}_{n_k} = (\mathbf{I} - 1/n\ 1_{n_k}1'_{n_k})$ matriz $(n_k \times n_k)$ $k = 0, 1, \ldots, r$.

Observe-se que $\mathbf{H}_{n_k}$ não é matriz de projeção pois ela é simétrica mas não é idempotente. ($\mathbf{H}^2 = \mathbf{I} - (n_k - 2)/n\ 1_{n_k}1'_{n_k} \neq \mathbf{H}$).

Agora, usando as matrizes **V** em (2.2.63) e **H** em (2.2.69) e aproveitando a simetria da matriz de dispersão, tem-se que S_v, dada em (2.2.68), pode ser escrita como segue, onde os elementos típicos dentro e entre os blocos, estão expressos de (a) a (i)

$$nS_v = \begin{array}{c} \begin{array}{cccccccc} z_1 & \cdots & z_r & y & z_1y & \cdots & z_ry & \end{array} \\ \left[\begin{array}{c|c|c|c|c|c|c} 1'_{n_1} H_{n_1} I_{n_1} & \cdots & \frac{-n_1 n_r}{n} & (a) & 1'_{n_1} H_{n_1} Y^{(1)} & \cdots & (b) \\ \hline & & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hline & & 1'_{n_r} H_{n_r} I_{n_r} & (c) & \frac{-n_r}{n} 1'_{n_1} Y^{(1)} & \cdots & (d) \\ \hline & & & (e) & (f) & \cdots & (g) \\ \hline & & & & Y^{(1)'} H_{n_1} Y^{(1)} & \cdots & (h) \\ \hline & & & & & \ddots & \vdots \\ \hline & & & & & & (i) \end{array}\right] \end{array} \begin{array}{c} z_1 \\ \vdots \\ z_r \\ y \\ z_1y \\ \vdots \\ z_ry \end{array}$$

(2.2.70)

onde

$$(a) = 1'_{n_1} H_{n_1} Y^{(1)} - \frac{n_1}{n} \sum_{\substack{m=0 \\ m \neq 1}}^{r} 1'_{n_m} Y^{(m)}$$

$$(b) = \frac{-n_1}{n} 1'_{n_r} Y^{(r)}$$

$$(c) = 1'_{n_r} H_{n_r} Y^{(r)} - \frac{n_r}{n} \sum_{\substack{m=0 \\ m \neq r}}^{r} 1'_{n_m} Y^{(m)}$$

$$(d) = 1'_{n_r} H_{n_r} Y^{(r)}$$

$$(e) = \sum_{k=0}^{r} \left[Y^{(k)'} H_{n_k} Y^{(k)} - \frac{1}{n} \sum_{\substack{m=0 \\ m \neq k}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)} \right]$$

$$(f) = Y^{(1)'} H_{n_1} Y^{(1)} - \frac{1}{n} \sum_{\substack{m=0 \\ m \neq 1}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_1} Y^{(1)}$$

$$(g) = Y^{(r)'} H_{n_r} Y^{(r)} - \frac{1}{n} \sum_{\substack{m=0 \\ m \neq r}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_r} Y^{(r)}$$

$$(h) = -\frac{1}{n} Y^{(1)'} 1_{n_1} 1'_{n_r} Y^{(r)}$$

$$(i) = Y^{(r)'} H_{n_r}^{(r)}$$

Agora, em S_v há alguns termos que podem ser escritos de maneira mais simples:

$$\widehat{\text{var}}(z_k) = \hat{\psi}^*_{kk} = \frac{1}{n} 1'_{n_k} H_{n_k} 1_{n_k} = \frac{1}{n} 1'_{n_k} (I - 1/n \; 1_{n_k} 1'_{n_k}) 1_{n_k}$$

$$= \frac{1}{n} 1'_{n_k} 1_{n_k} - \frac{1}{n^2} 1'_{n_k} 1_{n_k} 1'_{n_k} 1_{n_k} = \frac{n_k}{n} - \frac{n_k^2}{n^2}$$

então:

$$\hat{\psi}^*_{kk} = \frac{n_k}{n} \left(1 - \frac{n_k}{n} \right) \qquad k = 1, \ldots, r \qquad (2.2.71)$$

$$\widehat{\text{cov}}(z_k\ ,\ y) = (\hat{\delta}^*_{k1},\ \ldots,\ \hat{\delta}^*_{kp}) = \frac{1}{n}\, 1'_{n_k} H_{n_k} Y^{(k)} + \frac{-n_k}{n^2} \sum_{\substack{m=0 \\ m\neq k}}^{r} 1'_{n_m} Y^{(m)}$$

$$= \frac{1}{n}\, 1'_{n_k} Y^{(k)} - \frac{1}{n^2}\, 1'_{n_k} 1_{n_k} 1'_{n_k} 1_{n_k} Y^{(k)} \frac{n_k}{n^2} \sum_{\substack{m=0 \\ m\neq k}}^{r} \bar{y}^{(m)\prime} n_m$$

$$= \frac{n_k}{n}\, \bar{y}^{(k)\prime} - \frac{n_k^2}{n^2} \bar{y}^{(k)\prime} - \frac{n_k}{n^2} \sum_{\substack{m=0 \\ m\neq k}}^{r} n_m \bar{y}^{(m)\prime}$$

$$= \frac{n_k}{n}\, \bar{y}^{(k)\prime} - \frac{n_k}{n^2} \sum_{m=0}^{r} n_m \bar{y}^{(m)\prime}$$

portanto:

$$\widehat{\text{cov}}(z_k,\ y) = \frac{n_k}{n} (\bar{y}^{(k)} - \bar{y})' \qquad k = 1,\ \ldots,\ r \tag{2.2.72}$$

onde $\bar{y}^{(k)}$ e $\bar{y}$ foram definidos em (2.2.65) e (2.2.66).

$$\widehat{\text{cov}}(z_k, z_k y) = [(\hat{\varepsilon}_{11})_{k,k1},\ \ldots,\ (\hat{\varepsilon}_{11})_{k,kP}] = \frac{1}{n}\, 1'_{n_k} H_{n_k} Y^{(k)}$$

$$= \frac{n_k}{n}\, \bar{y}^{(k)\prime} - \frac{n_k^2}{n^2} \bar{y}^{(k)\prime}$$

$$= \frac{n_k}{n} \left(1 - \frac{n_k}{n} \right) \bar{y}^{(k)\prime} \qquad k = 1,\ \ldots,\ r \tag{2.2.73}$$

$$\widehat{\text{cov}}(z_i, z_k y) = [(\hat{\varepsilon}_{11})_{k,k1},\ \ldots,\ (\hat{\varepsilon}_{11})_{k,kp}] = \frac{-n_i}{n^2}\, 1'_{n_k} Y^{(k)}$$

$$= -\frac{n_i n_k}{n^2} \bar{y}^{(k)\prime} \qquad \begin{matrix} i \neq k \\ i,\ k = 1,\ \ldots,\ r \end{matrix} \tag{2.2.74}$$

$$\widehat{\text{cov}}\, y = \hat{\Gamma}^{*} = \frac{1}{n} \sum_{k=0}^{r} \left[Y^{(k)'} H_{n_k} Y^{(k)} - \frac{1}{n} \sum_{\substack{m=0 \\ m \neq k}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)} \right]$$

$$= \frac{1}{n} \sum_{k=0}^{r} \left[Y^{(k)'} Y^{(k)} - \frac{1}{n} Y^{(k)'} 1_{n_k} 1'_{n_k} Y^{(k)} - \frac{1}{n} \sum_{\substack{m=0 \\ m \neq k}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)} \right]$$

$$= \frac{1}{n} \sum_{k=0}^{r} \left[Y^{(k)'} Y^{(k)} - \frac{1}{n} \sum_{m=0}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)} \right]$$

$$= \frac{1}{n} \sum_{k=0}^{r} Y^{(k)'} Y^{(k)} - \frac{1}{n} \sum_{k=0}^{r} \left(\frac{1}{n} \sum_{m=0}^{r} n_m \bar{y}^{(m)} \right) n_k \bar{y}^{(k)'}$$

$$= \frac{1}{n} Y'Y - \bar{y} \left(\frac{1}{n} \sum_{k=0}^{r} n_k \bar{y}^{(k)'} \right) = \frac{1}{n} (Y'Y - n \bar{y} \bar{y}')$$

portanto:

$$\widehat{\text{cov}}\, y = \frac{1}{n} Y'H Y = S_y \qquad (2.2.75)$$

a matriz de covariância amostral de **y** é a usual obtida usando as n obsevações.

$$\widehat{\text{cov}}(y, z_k y) = \frac{1}{n} Y^{(k)'} H_{n_k} Y^{(k)} - \frac{1}{n^2} \sum_{\substack{m=0 \\ m \neq k}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)}$$

$$= \frac{1}{n} Y^{(k)'} Y^{(k)} - \frac{1}{n^2} \sum_{\substack{m=0 \\ m \neq k}}^{r} Y^{(m)'} 1_{n_m} 1'_{n_k} Y^{(k)}$$

$$= \frac{1}{n} \left(Y^{(k)'} Y^{(k)} - n_k \bar{y} \bar{y}^{(k)'} \right) \qquad k = 1, \ldots, r \qquad (2.2.76)$$

$$\widehat{\text{cov}}(z_i y,\ z_k y) = -\frac{1}{n^2}\, Y^{(k)'} 1_{n_i} 1'_{n_k} Y^{(k)} = -\frac{n_i n_k}{n^2}\, \bar{y}^{(i)} \bar{y}^{(k)'} \qquad (2.2.77)$$

$$i \neq k;\ i, k = 1, \ldots, r$$

Substituindo (2.2.71) a (2.2.77) em (2.2.70) a expressão da matriz de variância-covariância de **v** fica dada por

$$S_v = \begin{array}{c|cccccccc} & z_1 & \cdots & z_r & y & z_1 y & \cdots & z_r y & \\ \hline z_1 & \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right) & \cdots & \frac{-n_1 n_r}{n^2} & (a) & (b) & \cdots & (c) & z_1 \\ \hline \vdots & & \ddots & \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ \hline z_r & & & \frac{n_r}{n}\left(1 - \frac{n_r}{n}\right) & (d) & (e) & \cdots & (f) & z_r \\ \hline y & & & & (g) & (h) & \cdots & (i) & y \\ \hline z_1 y & & & & & (j) & \cdots & (k) & z_1 y \\ \hline \vdots & & & & & & \ddots & \vdots & \vdots \\ \hline z_r y & & & & & & & (l) & z_r y \end{array} \qquad (2.2.78)$$

onde,

$$(a) = \frac{n_1}{n}\left(\bar{y}^{(1)} - \bar{y}\right)'$$

$$(b) = \frac{n_1}{n}\left(1 - \frac{n_1}{n}\right)\bar{y}^{(1)'}$$

$$(c) = \frac{-n_1 n_r}{n^2}\ \bar{y}^{(r)'}$$

$$(d) = \frac{n_r}{n}\left(\bar{y}^{(r)} - \bar{y}\right)'$$

$$(e) = \frac{-n_1 n_r}{n^2}\ \bar{y}^{(1)'}$$

$$(f) = \frac{n_r}{n}\left(1 - \frac{n_r}{n}\right)\bar{y}^{(r)'}$$

$$(g) = S_y = \frac{1}{n}\,Y'H_nY = \frac{1}{n}\sum_{k=0}^{r} Y^{(k)'}\tilde{H}_{n_k}Y^{(k)} + \sum_{k=0}^{r}\frac{n_k}{n}(\bar{y}^{(k)}-\bar{y})(\bar{y}^{(k)}-\bar{y})'$$

$$(h) = \frac{1}{n}\left(Y^{(1)'}Y^{(1)} - n_1\bar{y}\,\bar{y}^{(1)'}\right)$$

$$(i) = \frac{1}{n}\left(Y^{(r)'}Y^{(r)} - n_r\bar{y}\,\bar{y}^{(r)'}\right)$$

$$(j) = \frac{1}{n}\,Y^{(1)'}H_{n_1}Y^{(1)}$$

$$(k) = \frac{-n_1 n_r}{n^2}\ \bar{y}^{(1)}\bar{y}^{(r)'}$$

$$(l) = \frac{1}{n}\,Y^{(r)'}H_{n_r}Y^{(r)}$$

onde, em (g) $\tilde{H}_{n_k} = I - \frac{1}{n_k}1_{n_k}1'_{n_k}$.

c) Misturas de multinomial e contínuas.

Para terminar apresentam-se os estimadores da média e matriz de dispersão do vetor $w^* = (z', y')'$, obtidos truncando convenientemente os estimadores $\bar{v}$ e S_v, para o vetor

$v = (w^{*'}, (z \otimes y)')'$, dados em (2.2.67) e (2.2.70) respectivamente.

Seja a matriz de dados:

$$W^{*} = \begin{bmatrix} z_{11} & \cdots & z_{1r} & y_{11} & \cdots & y_{1p} \\ z_{21} & \cdots & z_{2r} & y_{21} & \cdots & y_{2p} \\ \vdots & & \vdots & \vdots & & \vdots \\ z_{n1} & \cdots & z_{nr} & y_{n1} & \cdots & y_{np} \end{bmatrix} \tag{2.2.79}$$

As observações de W^{*} podem ser reordenadas tal como foi feito com a matriz V. Logo, W^{*} pode ser reescrita em forma de blocos de maneira análoga a V em (2.2.63); assim:

$$\begin{array}{c} \\ n_1 \\ \vdots \\ n_r \\ n_0 \end{array} \begin{array}{c} \begin{array}{cccc} z_1 & z_2 & \cdots & z_r \end{array} \\ \left[\begin{array}{c|c|c|c|c} 1_{n_1} & O_{n_1} & \cdots & O_{n_1} & Y^{(1)} \\ \hline \vdots & \vdots & \cdots & \vdots & \vdots \\ \hline O_{n_r} & O_{n_r} & \cdots & 1_{n_r} & Y^{(r)} \\ \hline O_{n_0} & O_{n_0} & \cdots & O_{n_0} & Y^{(0)} \end{array}\right] \end{array} \tag{2.2.80}$$

Com esta matriz calcula-se a média e variância amostral para w^{*} e tem-se que, de (2.2.68),

$$\bar{w}^{*} = \left(\frac{n_1}{n}, \frac{n_2}{n}, \ \ldots, \frac{n_r}{n}, \bar{y}' \right)' = \left(\frac{n_1}{n}, \ \ldots, \frac{n_1}{n}, \ \bar{y}_1, \ \ldots, \ \bar{y}_p \right)' \tag{2.2.81}$$

agora, de (2.2.78) segue

$$
S_w^* = \begin{array}{c}
\begin{array}{ccccc} z_1 & & \cdots & & z_r \end{array} \qquad\qquad y \\
\left[\begin{array}{c|c|c|c}
\dfrac{n_1}{n}\left(1 - \dfrac{n_1}{n}\right) & \cdots & \dfrac{-n_1 n_r}{n^2} & \dfrac{n_1}{n}\left(\bar{y}^{(1)} - \bar{y}\right)' \\
\hline
\vdots & \ddots & \vdots & \vdots \\
\hline
\dfrac{-n_r n_1}{n^2} & \cdots & \dfrac{n_r}{n}\left(1 - \dfrac{n_r}{n}\right) & \dfrac{n_r}{n}\left(\bar{y}^{(r)} - \bar{y}\right)' \\
\hline
\dfrac{n_1}{n}\left(\bar{y}^{(1)} - \bar{y}\right) & \cdots & \dfrac{n_r}{n}\left(\bar{y}^{(r)} - \bar{y}\right) & \dfrac{1}{n}\, Y'\, H\, Y
\end{array}\right]
\begin{array}{c} z_1 \\ \vdots \\ z_r \\ \\ y \end{array}
\end{array}
$$

(2.2.82)

CAPÍTULO 3

MÉTODOS DE DISCRIMINAÇÃO PARA MISTURAS

3.1 - INTRODUÇÃO

Entre os métodos de discriminação desenvolvidos especificamente para tratar problemas com misturas de variáveis, é possível identificar duas linhas definidas de evolução que serão os principais focos de atenção deste trabalho.

A primeira linha de evolução está baseada no modelo de posição, proposto por Olkin e Tate (1961) e extendido por Afifi e Elashoff (1969). Este modelo foi usado por primeira vez em problemas de classificação por Chang e Afifi (1974), eles sugeriram um método adequado para discriminar com p variáveis contínuas e uma binária. Em 1975, Krzanowski propôs uma extensão para o caso de p variáveis contínuas e q binárias ($q \geq 1$). Já em 1980, o mesmo autor adaptou o método para tratar misturas de variáveis contínuas e categóricas em

geral (não somente binárias). Todo o desenvolvimento do método de discriminação baseado no modelo de posição é descrito em detalhe na seção 3.2.

Na seção 3.3 estuda-se a segunda linha de evolução. Esta aparece sobre a construção de funções discriminantes lineares incorporando as variáveis categóricas (ou modificações delas), no papel de variáveis independentes. Primeiramente, apresenta-se a adaptação da FLD de Fisher para misturas de Krzanowski (1975), e depois as modificações propostas por Vlachonikolis e Marriott (1982).

Além dos métodos mencionados acima, existem outros apropriados para discriminar com misturas. De um lado está o procedimento de discriminação baseado na regressão logística proposto por Cornfield (1962), Cox (1966) e Day e Kerridge (1967) para certas populações não normais (entres as quais estão as populações caracterizadas por vetores mistos de variáveis contínuas e binárias). De outro lado, estão a discriminação Kernel e a discriminação segundo o vizinho mais próximo e suas adaptações para tratar misturas. A discriminação logística e os outros métodos são tratados nas seções 3.4 e 3.5 respectivamente.

3.2 - DISCRIMINAÇÃO BASEADA NO MODELO DE POSIÇÃO

Seguindo um enfoque estritamente paramétrico, o primeiro passo antes de realizar qualquer análise de misturas de variáveis, deveria ser a formulação de um modelo apropriado que sirva de base à análise. A distribuição normal multivariada é geralmente usada nas análises com dados contínuos enquanto que a distribuição multinomial tem mostrado ser a base natural das análises com dados categóricos. Isto leva a pensar que uma combinação destas distribuições poderia fornecer um modelo apropriado para mistura de variáveis. Um modelo satisfatório,

seria aquele que especificasse a distribuição conjunta de todas as variáveis de forma tal que, a associação entre variáveis categóricas e contínuas e as características marginais de cada tipo de variável, possam ser modeladas.

Intuitivamente, a primeira coisa a fazer para encontrar um modelo adequado é dividir as variáveis em dois conjuntos, pois a distribuição conjunta delas pode ser escrita como o produto da distribuição marginal de algumas variáveis pela distribuição condicional das restantes dados os valores das primeiras. A divisão natural é em variáveis contínuas e categóricas. Olkin e Tate (1961) propuseram um modelo no qual a distribuição marginal das variáveis categóricas está multiplicada pela distribuição condicional das variáveis contínuas dado algum valor das categóricas. Este modelo é conhecido na literatura como *Location Model* que é traduzido como modelo de posição.

3.2.1 - O Modelo de Posição

No modelo de posição assume-se que as variáveis categóricas estão arranjadas em forma de tabela de contingência, assim, cada padrão dos valores destas variáveis determina de maneira única uma categoria ou posição na tabela. Supõe-se também que a distribuição condicional das variáveis contínuas é normal multivariada, cujos parâmetros dependem da posição na tabela, ou seja, das variáveis categóricas. Observe-se que, se o vetor de variáveis categóricas é binário, este pode ser transformado em multinomial usando a definição 2.5.1.

Por outro lado o vetor contínuo $\mathbf{y} = (y_1, \ldots, y_p)'$ segue uma distribuição normal multivariada com média $\mu^{(m)}$ na categoria m, e matriz de dispersão Σ, comum em todas as categorias. Por último, assume-se que a probabilidade de obter uma observação da categoria m é p_m.

Originalmente o modelo de posição foi proposto para estudar a correlação multivariada entre vetores mistos de variáveis multinomiais e contínuas em uma população (ver Olkin e Tate, 1961). Posteriormente foi extendido por Afifi e Elashoff (1969) para testar a hipótese nula, conhecida como 'hipótese de posição' (*location hypotesis*), de que os parâmetros correspondetes às d categorias mutinomiais, (p_1, ..., p_d, $\mu^{(1)}$, ..., $\mu^{(d)}$), são iguais em duas populações. Seguindo essa generalização para tratar o problema de classificação de interesse, Krzanowski (1975) assume que **y** tem uma distribuição normal p-variada com média $\mu_i^{(m)}$ na m-ésima categoria da população π_i (i = 1, 2) e matriz de covariância Σ comum em todas as categorias e nas duas populações, isto é,

$$(\mathbf{y} \mid z_m = 1,\ z_j = 0\ j \neq m = 1, \ldots, d) \sim N_p(\mu_i^{(m)}, \Sigma) \quad \text{em } \pi_i\ i = 1, 2 \tag{3.2.1}$$

Além disso, a probabilidade de obter uma observação na categoria m da população π_i é p_{im} (i = 1, 2; m = 1, ..., d).

É interessante notar que apesar das matrizes de dispersão condicional de $\mathbf{y}|\mathbf{x}$ serem iguais em π_1 e π_2, as covariâncias marginais de **y** são diferentes, pois dependem das probabilidades multinomiais p_{im} (ver (2.2.30)). Portanto, conclui-se que as matrizes de covariância de **w** também variam de uma população à outra.

Uma vez que o modelo já foi apresentado, a regra de discriminação ótima será constituída, seguindo a teoria geral de classificação (ver Anderson, 1958 Cap. 6).

3.2.2 - Regra de Classificação para Misturas de Variáves Binárias e Contínuas

Nesta seção será tratado o vetor misto $\mathbf{w} = (x_1, \ldots, x_q, y_1, \ldots, y_p)'$ de variáveis binárias e contínuas, definido na parte (a) da seção 2.2.1. Como já foi visto na seção 3.2.1, no modelo de posição assume-se que as variáveis categóricas são multinomiais, é por esta razão que antes de começar a análise, o vetor $\mathbf{w}$ deve ser transformado em um vetor misto de variáveis multinomiais e contínuas usando a definição 2.1.6. Assim, de $\mathbf{w} = (x_1, \ldots, x_q, y_1, \ldots, y_p)'$ passa-se a $\mathbf{w} = (z_1, \ldots, z_{2^q}, y_1, \ldots, y_p)'$ com

$$\mathbf{z}' = (z_1, \ldots, z_{2^q}) \sim \text{Multin}(1, p_{i1}, \ldots, p_{i2^q}) \text{ em } \pi_i, \; i = 1, 2.$$

Para começar, a regra de classificação será construída para o caso de parâmetros conhecidos.

Seja $f_i(\mathbf{w})$ a função de distribuição de probabilidades de $\mathbf{w}$ em π_i ($i = 1, 2$). É fácil demonstrar que a regra ótima no sentido Bayesiano (supondo custos e probabilidades prévias iguais) é:

$$\text{Alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } \frac{f_1(\mathbf{w})}{f_2(\mathbf{w})} \geq 1 \qquad (3.2.2)$$

$$\text{e em } \pi_2 \text{ caso contrário.}$$

Agora:

$$f_i(\mathbf{w}) = f_i(\mathbf{x}',\mathbf{y}')' = f_i(\mathbf{z}',\mathbf{y}')' = f_i(\mathbf{z})\, f_i(\mathbf{y} \mid \mathbf{z}) \qquad i = 1, 2$$

então, se condicionada em $\mathbf{x}$ e supondo que a observação $\mathbf{w}$ pertence à m-ésima categoria, $f_i(\mathbf{w})$ pode ser escrita da seguinte maneira:

$$f_i(\mathbf{w}) = p_{im} f_i(\mathbf{y} \mid \mathbf{z}_m = 1) \qquad i = 1, 2; \; m = 1, \ldots, 2^q. \qquad (3.2.3)$$

Usando (3.2.1) e (3.2.3) em (3.2.2) pode-se construir a seguinte

regra de classificação para uma observação mista $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$:

$$\text{Alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } \frac{p_{1m}\exp\{-\frac{1}{2}(\mathbf{y}-\mu_1^{(m)})'\,\Sigma^{-1}(\mathbf{y}-\mu_1^{(m)})\}}{p_{2m}\exp\{-\frac{1}{2}(\mathbf{y}-\mu_2^{(m)})'\,\Sigma^{-1}(\mathbf{y}-\mu_2^{(m)})\}} \geq 1$$

e em π caso contrário.

Aplicando logarítmo ao termo à esquerda tem-se:

$$\ln\left(\frac{p_{1m}}{p_{2m}}\right) - \frac{1}{2}(\mathbf{y}-\mu_1^{(m)})'\Sigma^{-1}(\mathbf{y}-\mu_1^{(m)}) + \frac{1}{2}(\mathbf{y}-\mu_2^{(m)})'\Sigma^{-1}(\mathbf{y}-\mu_2^{(m)})$$

simplificando esta expressão, a regra de classificação pode ser escrita como:

Dado $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$, se $m = 1 + \sum_{l=1}^{q} x_l 2^{(l-1)}$ então,

$$\text{alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } (\mathbf{y} - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})'\Sigma^{-1}(\mu_1^{(m)} - \mu_2^{(m)}) \geq \ln\left(\frac{p_{2m}}{p_{1m}}\right)$$

e em π_2 caso contrário. (3.2.4)

Observe-se que a regra ótima derivada do modelo de posição, gera uma função linear discriminante diferente para cada categoria multinomial e que os pontos de corte correspondentes, dependem unicamente das componentes discretas do modelo. Note-se também, que as funções de discriminação representam hiperplanos distintos cujos coeficientes dependem dos vetores de médias associado a cada categoria e da matriz de dispersão.

Chang e Afifi (1974), trataram o caso de discriminação com uma variável binária e p contínuas a partir de um modelo bisserial pontual. Este modelo foi estudado em detalhe por Tate (1954) para o caso em que **x** é binomial e a distribuição condicional de **y** dado **x** é

normal. O modelo proposto por Olkin e Tate (1961) surgiu como uma extensão multivariada do modelo bisserial pontual, na qual $\mathbf{x} = (x_1, \ldots, x_k)$ tem distribuição multinomial e a distribuição condicional de $\mathbf{y} = (y_1, \ldots, y_p)$ para $\mathbf{x}$ fixo é normal multivariada.

Chang e Afifi no mesmo artigo derivaram a solução Bayesiana para esse problema assumindo médias e matrizes de variância e covariância modeladas por:

$$\mu_{1x} = \mu_1 + x\Delta_1 \qquad x = 0, 1 \qquad i = 1, 2$$

e

$$\Sigma_x = \Sigma + x\Gamma \qquad x = 0, 1$$

assim, a regra obtida por eles depende de duas funções de discriminação, uma para cada valor de x, que foi chamada pelos autores de Função Discriminante Dupla (*Double Discriminant Function*).

Continuando com o modelo de posição tem-se que, para calcular as probabilidades de má classificação é necessário conhecer a distribuição da função linear, $\xi(m)$ dada por:

$$\xi(m) = (\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1}(y - \frac{1}{2}(\mu_1^{(m)} + \mu_2^{(m)})) \qquad m = 1, \ldots, 2^q.$$

Note-se que para cada categoria m, $\xi(m)$ é uma transformação linear de y e portanto, a sua distribuição também é normal (ver Mardia *et al.*, 1979 p. 61). Na população π_i, os parâmetros são os seguintes:

$$E(\xi(m)) = (\mu_1^{(m)} - \mu_2^{(m)})' \Sigma^{-1}\mu_1^{(m)} - \frac{1}{2}(\mu_1^{(m)} - \mu_2^{(m)})'\Sigma^{-1}(\mu_1^{(m)} + \mu_2^{(m)})$$

$$Var(\xi(m)) = (\mu_1^{(m)} - \mu_2^{(m)})'\Sigma^{-1}(\mu_1^{(m)} - \mu_2^{(m)}) = D_m^2$$

onde D_m^2 é a distância quadrada de Mahalanobis entre π_1 e π_2, condicionada a que a observação pertença à m-ésima categoria.

Assim,

$$\text{se } \mathbf{w} \in \pi_1 \quad E(\xi(m)) = \frac{1}{2}(\mu_1^{(m)}-\mu_2^{(m)})'\Sigma^{-1}(\mu_1^{(m)}-\mu_2^{(m)}) = \frac{1}{2}D_m^2$$

e

$$\text{se } \mathbf{w} \in \pi_2 \quad E(\xi(m)) = -\frac{1}{2}(\mu_1^{(m)}-\mu_2^{(m)})'\Sigma^{-1}(\mu_1^{(m)}-\mu_2^{(m)}) = -\frac{1}{2}D_m^2$$

então,

$$\text{se } \mathbf{w} \in \pi_1 \quad \left(\frac{\xi(m) - \frac{1}{2}D_m^2}{D_m} \right) \sim N(0,1)$$

e

$$\text{se } \mathbf{w} \in \pi_2 \quad \left(\frac{\xi(m) + \frac{1}{2}D_m^2}{D_m} \right) \sim N(0,1)$$

portanto, as probabilidades de classificação errada são

$$P(2|1) = \sum_{m=1}^{2^q} p_{1m} \, P[\xi(m) < \ln (p_{2m}/p_{1m})]$$

$$= \sum_{m=1}^{2^q} p_{1m} \, \Phi \left(\frac{\ln (p_{2m}/p_{1m}) - \frac{1}{2}D_m^2}{D_m} \right)$$

e

$$P(1|2) = \sum_{m=1}^{2^q} p_{2m} \, P[\xi(m) \geq \ln (p_{2m}/p_{1m})]$$

$$= \sum_{m=1}^{2^q} p_{2m} \, \Phi \left(\frac{\ln (p_{1m}/p_{2m}) - \frac{1}{2}D_m^2}{D_m} \right)$$

onde,

$P(i \mid j)$ representa a probabilidade de classificar uma observação em π_i dado que pertence a π_j $(i \neq j,\ i, j = 1, 2)$ e

Φ representa a função de distribuição acumulada da normal padrão.

Na prática, geralmente não é possível usar a regra populacional pois os parâmetros quase nunca são conhecidos, a informação disponível, provém de duas amostras de treinamento de tamanhos n_1 e n_2, de π_1 e π_2 respectivamente, e é com elas que os parâmetros são estimados e depois construída a regra amostral.

Para simplificar o processo de estimação, as frequências de ocorrência de cada padrão de $\mathbf{x}$ são escritas na forma de tabela de contingência com 2^q categorias assim, n_{im} é a frequência da categoria m na população π_i ($i = 1, 2$; $m = 1, \ldots, 2^q$). Agora, se $y_{ij}^{(m)}$ representa o vetor de variáveis contínuas associadas à j-ésima observação da categoria m da população π_i então, os estimadores de máxima verossimilhança de p_{im} e $\mu_i^{(m)}$ são respectivamente,

$$\hat{p}_{im} = \frac{n_{im}}{n_i} \tag{3.2.5}$$

$$\hat{\mu}_i^{(m)} = \bar{y}_i^{(m)} = \frac{1}{n_{im}} \sum_{j=1}^{n_{im}} y_{ij}^{(m)} \tag{3.2.6}$$

enquanto que o estimador não viciado de Σ está dado por,

$$\hat{\Sigma} = \frac{1}{n_1 + n_2 - 2^{q+1}} \sum_{i=1}^{2} \sum_{m=1}^{2^q} \sum_{j=1}^{n_{im}} (y_{ij}^{(m)} - \bar{y}_i^{(m)})(y_{ij}^{(m)} - \bar{y}_i^{(m)})', \tag{3.2.7}$$

$$i = 1, 2; \; m = 1, 2, \ldots, 2^q.$$

Logo, a regra de classificação estimada é:

Dado $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$, se $m = 1 + \sum_{l=1}^{q} x_l 2^{(l-1)}$ então,

alocar w em π_1 se

$$(\bar{y}_1^{(m)} - \bar{y}_2^{(m)})'\hat{\Sigma}^{-1}(y - \frac{1}{2}(\bar{y}_1^{(m)} + \bar{y}_2^{(m)})) \geq \ln\left(\frac{\hat{p}_{2m}}{\hat{p}_{1m}}\right) \qquad (3.2.8)$$

e em π_2 caso contrário.

Os estimadores anteriores (3.2.5) a (3.2.7) são úteis somente quando n_1 e n_2 são grandes em relação a 2^q, se não for assim, é muito provável que algum n_{im} seja pequeno ou zero e, em consequência, as estimativas dos parâmetros para estas categorias serão pobres ou não será possível obtê-las. Este problema leva a pensar que para que a regra (3.2.8) seja útil, deve-se encontrar alguma outra forma de obter bons estimadores para os parâmetros de todas as categorias multinomiais. Krzanowski (1975) propôs impor uma estrutura adicional à distribuição das variáveis para que os parâmetros $\mu_i^{(m)}$ sejam estimados a partir de um modelo linear e as probabilidades de ocorrência de cada categoria em cada grupo, p_{im}, sejam estimadas usando um modelo log-linear. Em detalhe, a proposta é a seguinte:

- Estimação de p_{im}

Considere-se primeiro as variáveis binárias separadamente. Usando a informação contida nelas, o autor escreve as amostras na forma de tabelas de contingência e depois usa modelos log-lineares para analisá-las. Nestes modelos, o logarítmo da frequência observada na m-ésima categoria de π_i, n_{im}, pode ser expressada como combinação linear dos efeitos principais e as interações de toda ordem (ver Goldstein e Dillon, 1978, p. 23). Krzanowski (1975) assume que n_{im} é uma realização de uma variável de média não nula η_{im}, que satisfaz uma relação da forma:

$$\log \eta_{im} = \sum_j a_{imj}\theta_j \qquad (3.2.9)$$

onde as a_{imj}'s representam constantes conhecidas e os θ_j's são um

conjunto de parâmetros desconhecidos que representam os efeitos principais das variáveis binárias e todas as possíveis interações entre elas. A estimação pelo método de máxima verossimilhança destes parâmetros, não é direta e pode ser feita usando o procedimento de ajuste proporcional iterativo (*iterative proportional fitting*), descrito por exemplo, em Deming e Stephan (1940) ou Fienberg (1970) (ver também Haberman, (1972) para o algorítmo do método e a programação computacional).

Observe-se que se o modelo completo (3.2.9) fosse ajustado, (i.e., se fossem consideradas todas as interações até ordem q) as frequências observadas seriam recuperadas e isto não resolveria o problema. Uma solução imediata é procurar um modelo reduzido que forneça estimativas não nulas de todas as frequências esperadas.

Krzanowski (1975) afirma que uma aproximação adequada para muitos dos problemas práticos é a aproximação de segunda ordem, que consiste em reter no modelo somente os efeitos principais e as interações de primeira ordem.

Se as tabelas de contingência são muito esparsas (*sparse*) pode acontecer que, embora seja usada a aproximação anterior, ainda restem algumas categorias nas quais $\hat{\eta}_{im} = 0$. Uma solução prática é omitir no modelo ajustado os termos de interação e trabalhar só com os efeitos principais.

Uma vez que as frequências esperadas são estimadas, os estimadores dos parâmetros p_{im} estão dados por:

$$\hat{p}_{im} = \frac{\hat{\eta}_{im}}{n_1} \qquad i = 1, 2 \qquad m = 1, \ldots, 2^q.$$

- Estimação de $\mu_i^{(m)}$ e Σ

Considere-se agora os parâmetros $\mu_i^{(m)}$ e Σ, relacionados às variáveis contínuas. Segundo Krzanowski (1975), uma estrutura bem simples e também compatível com o modelo de posição é o modelo linear aditivo cujos componentes representem efeitos principais e todas as possíveis interações entre as variáveis binárias assim, a média de **y** em π_i (i = 1, 2) pode ser escrita como:

$$\mu_i = \nu_i + \sum_{j=1}^{q} \alpha_{i,j} x_j + \sum_{j} \sum_{k>j} \beta_{i,jk} x_j x_k + \ldots + \delta_{i,1\ldots q} x_1 \ldots x_q \qquad (3.2.10)$$

ou na forma matricial

$$\mu_i = \beta_i^{*\prime} \upsilon^* \qquad (3.2.11)$$

onde, $\upsilon^* = (1, x_1, \ldots, x_q, x_1x_2, \ldots, x_1x_2\ldots x_q)'$

e $\beta_i^{*\prime} = \left[\nu_1 | \alpha_{i1} | \alpha_{i2} | \ldots | \beta_{i12} | \ldots | \delta_{i12\ldots q} \right]_{px2^q}$.

As médias condicionais $\mu_i^{(m)}$ são obtidas inserindo os valores das variáveis binárias, correspondentes à m-ésima categoria, no lado direito da expressão (3.2.10).

Novamente, neste caso, é melhor trabalhar com um modelo aproximado tratando os termos de maior ordem como erro residual e estimando os parâmetros que ficaram no modelo usando regressão multivariada (ver Mardia *et al.*, 1979 Cap. 6). Krzanowski (1975), sugere que por razões de consistência, a ordem do modelo linear ajustado seja igual à ordem do modelo log-linear usado para estimar as probabilidades multinomiais. Assim, para um modelo linear de segunda ordem, basta considerar os parâmetros ν_i, α_{ij}, β_{ijk} em (3.2.10), logo

dados,

$$\upsilon = (1, x_1, x_2, \ldots, x_q, x_1x_2, x_1x_3, \ldots, x_{q-1}x_q)'$$

e

$$B_i' = [\nu_i | \alpha_{i1} | \alpha_{i2} | \ldots | \alpha_{iq} | \beta_{i12} | \ldots | \beta_{iq-1q}]_{pxt}$$

onde B_i é uma matriz com pxt, $t = 1 + q(q + 1)/2$, parâmetros a serem estimados tem-se que, o modelo linear reduzido para estimar a média condicional do vetor contínuo y em π_i é:

$$y = B_i'\upsilon + \varepsilon \qquad \text{com} \quad \varepsilon \sim N_p(0, \Sigma) \tag{3.2.12}$$

agora, como dado x, o valor do vetor υ é conhecido, tem-se que:

$$y|x \sim N_p(B_i'\upsilon, \Sigma) \quad \text{em } \pi_i \qquad i = 1, 2.$$

Para o caso amostral, supondo ter n_i observações da população π_i ($i = 1, 2$), o modelo de regressão multivariada para estimar os parâmetros está dado por:

$$Y_i = X_iB_i + \varepsilon_i \tag{3.2.13}$$

onde,

$$Y_i = \begin{bmatrix} y_{11} & \cdots & y_{1p} \\ \vdots & & \vdots \\ y_{n_i1} & \cdots & y_{n_ip} \end{bmatrix}_{n_i \times p}$$

$$X_i = \begin{bmatrix} 1 & x_1^{(1)} & \cdots & (x_{q-1}x_q)^{(1)} \\ \vdots & & & \vdots \\ 1 & x_1^{(ni)} & \cdots & (x_{q-1}x_q)^{(ni)} \end{bmatrix}_{n_i xt}$$

$$\mathbf{B}_i = \begin{bmatrix} \nu_i^{(1)} & \cdots & \nu_i^{(p)} \\ \alpha_{i1}^{(1)} & \cdots & \alpha_{i1}^{(p)} \\ \vdots & & \vdots \\ \beta_{i,q-1\,q}^{(1)} & \cdots & \beta_{i,q-1\,q}^{(p)} \end{bmatrix}_{txp} = \begin{bmatrix} \nu_i' \\ \hline \alpha_{i1}' \\ \hline \vdots \\ \hline \beta_{i,q-1\,q}' \end{bmatrix}$$

$$\varepsilon_i = [e_{ij}]_{n_i \times p}$$

Agora, seguindo Mardia *et al.* (1979, p. 158), os estimadores de máxima verossimilhança de $\mathbf{B}$ e Σ, em π_i, quando $n_i \geq p + t$ e $\mathbf{X}_i$ é de posto coluna completo t, são:

$$(1)\ \hat{\mathbf{B}}_i = (\mathbf{X}_i' \mathbf{X}_i)^{-1} \mathbf{X}_i' \mathbf{Y}_i \qquad (3.2.14)$$

$$\hat{\Sigma}_i = \frac{1}{n_i} \mathbf{Y}_i' \mathbf{P}_i \mathbf{Y}_i \qquad (3.2.15)$$

onde, $\mathbf{P}_i = \mathbf{I} - \mathbf{X}_i(\mathbf{X}_i'\mathbf{X}_i)^{-1}\mathbf{X}_i'$.

Se as condições anteriores não são satisfeitas ou se o modelo ajustado para as variáveis binárias for de primeira ordem, os produtos $x_i x_j$ e os parâmetros β_{ijk} podem ser omitidos em $\mathbf{X}_i$ e $\mathbf{B}_i$, com esta aproximação espera-se obter os estimadores necessários. Outro detalhe a ser observado é que o estimador (3.2.15) não pode ser usado diretamente pois uma das suposições do problema em estudo é que a matriz de dispersão é comum às duas populações, logo, o estimador de máxima verossimilhança de Σ é uma média ponderada de $\hat{\Sigma}_1$ e $\hat{\Sigma}_2$ (ver Mardia *et al.* p.140),

---

(1) $\hat{\mathbf{B}}_i$ está bem definido pois a condiçao $\text{posto}(\mathbf{X}_i) = t$ implica em que a inversa $(\mathbf{X}_i'\mathbf{X}_i)^{-1}$ existe.

$$\hat{\Sigma} = \frac{n_1 \hat{\Sigma}_1 + n_2 \hat{\Sigma}_2}{n_1 + n_2}$$

ou equivalentemente:

$$(n_1 + n_2)\,\hat{\Sigma} = Y_1' P_1 Y_1 + Y_2' P_2 Y_2 \qquad (3.2.16)$$

Agora, o estimador da média condicional de y na m-ésima categoria é

$$\bar{y}^{(m)} = \hat{B}_1' v^{(m)} \qquad (3.2.17)$$

onde $v^{(m)}$ é o valor do vetor v na m-ésima categoria multinomial (m = 1, 2, ..., 2^q).

A partir dos estimadores apropriados, baseados num modelo restrito, de todos os parâmetros necessários é possível construir a regra de classificação estimatida dada em (3.2.8).

3.2.3 - Extensão do Método para Misturas de Variáveis Categóricas e Contínuas

Em 1980, Krzanowski propôs uma modificação muito simples do procedimento anterior a qual permite classificar observações de populações caracterizadas por vetores mistos de variáveis contínuas e variáveis categóricas de qualquer tipo. Esta modificação, que está baseada no fato de que qualquer variável categórica pode ser representada por um conjunto de variáveis binárias, será apresentada nesta seção.

Seja o vetor misto $w = (x_1, \ldots, x_c, y_1, \ldots, y_p)'$ formado por p

variáveis contínuas $y_1, \ldots, y_p$ e c variáveis categóricas $x_1, \ldots, x_c$ tais que a i-ésima tem k_i categorias (i = 1, ..., c). (É claro que as variáveis binárias podem ser consideradas como categóricas de duas categorias).

As c variáveis categóricas de **w** podem ser combinadas em uma única multinomial com $k = \prod_{i=1}^{c} k_i$ categorias, usando a definição 2.1.5. Logo, o modelo de posição associado a **w** será o mesmo que o anterior com a distribuição condicional de y na categoria m de π_i, $N_p(\mu_i^{(m)}, \Sigma)$ e a probabilidade de pertencer à m-ésima categoria de π_i igual a p_{im} (i = 1, 2; m = 1, ..., k).

Se o número total de categorias geradas pelas variáveis categóricas é grande, o número de parâmetros a ser estimados também crescerá muito, portanto neste caso também será necessário modelar as médias condicionais de y e as probabilidades multinomiais p_{im}, com a finalidade de obter as estimativas necessárias para construir a regra de classificação.

Como no caso de q variáveis binárias, as probabilidades de ocorrência das categorias são modeladas com um modelo log-linear e são estimadas a partir de tabelas de contingência construidas considerando as k categorias e usando o método iterativo proporcional descrito em Haberman (1972). Apesar deste método fornecer as estimativas das p_{im}'s, alguns problemas surgem, pois não é tão simples definir exatamente o significado de um modelo de primeira ou segunda ordem quando algumas das variáveis categóricas têm mais de duas categorias. Esta situação fica mais complicada quando se define diretamente um modelo adequado para estimar $\mu_i^{(m)}$ e Σ através de regressão multivariada. O seguinte exemplo mostrará algumas das dificuldades.

Exemplo 3.2.1: Sejam 3 variáveis categóricas,

x_1 com 2 categorias que assumem valores: 0 e 1;

x_2 com 3 categorias que assumem valores: 0, 1 e 2;

x_3 com 3 categorias que assumem valores: 1, 2 e 3.

O modelo de regressão linear de segunda ordem para estimar as médias condicionais e a matriz de dispersão (sem considerar ainda as populações) é:

$$\mu^{(m)} = \nu + \alpha_1 x_1 + \alpha_2 x_2 + \alpha_3 x_3 + \beta_{12} x_1 x_2 + \beta_{13} x_1 x_3 + \beta_{23} x_2 x_3 + \varepsilon.$$

A seguir serão comparados três categorias nas quais só muda o valor observado na terceira variável.

Seja

x	m	$\mu^{(m)}$
(1,2,1)	1	$\nu + \alpha_1 + 2\alpha_2 + \alpha_3 + 2\beta_{12} + \beta_{13} + 2\beta_{23}$
(1,2,2)	2	$\nu + \alpha_1 + 2\alpha_2 + 2\alpha_3 + 2\beta_{12} + 2\beta_{13} + 4\beta_{23}$
(1,2,3)	3	$\nu + \alpha_1 + 2\alpha_2 + 3\alpha_3 + 2\beta_{12} + 3\beta_{13} + 6\beta_{23}$

então:

$$\mu^{(2)} - \mu^{(1)} = \alpha_3 + \beta_{13} + 2\,\beta_{23}$$
$$\mu^{(3)} - \mu^{(2)} = \alpha_3 + \beta_{13} + 2\,\beta_{23}$$

logo, pode-se escrever:

$$\mu^{(2)} = \mu^{(1)} + (\alpha_3 + \beta_{13} + 2\,\beta_{23})$$
$$\mu^{(3)} = \mu^{(1)} + 2(\alpha_3 + \beta_{13} + 2\,\beta_{23})$$

portanto, o modelo usado assume que existe relação linear entre as

categorias das variáveis e isto não faz nenhum sentido se as variáves envolvidas são nominais. Agora, se a variável x_3 fosse ordinal o modelo do exemplo supõe equidistância entre categorias consecutivas e isto não é necessariamente verdade. ■

Os problemas de interpretação dos modelos lineares e log-lineares que aparecem quando são consideradas variáveis de mais de duas categorias podem ser evitados usando um artifício simples e muito conhecido que permite passar da nova situação à anterior. Trata-se de definir variáveis *dummy* para representar cada variável categórica por um conjuto de variáveis binárias. Esta transformação foi apresentada na definição 2.1.4.

Depois de modificar as variáveis categóricas convenientemente o problema original foi reduzido a outro de variáveis binárias e contínuas que pode ser resolvido usando o procedimento descrito na seção anterior. Porém, basta refletir um pouco para notar que, neste caso, a aplicação mecânica e direta daquele procedimento implica em um 'superdimensionamento' do problema de estimação ou, dito de outra forma, em um aumento irreal do número de parâmetros a serem estimados.

Considere primeiro, o modelo linear de segunda ordem para estimar as médias condicionais $\mu_i^{(m)}$. Este modelo inclui todos os efeitos principais das variáveis binárias e as interações de primeira ordem repesentadas pelos produtos cruzados dessas variáveis duas a duas. Porém, quando uma variável categórica é representada por um conjunto de variáveis binárias, somente uma destas pode ser não nula, logo não faz sentido incluir no modelo de regressão multivariada os termos de interação entre as variáveis binárias que representam a mesma variável categórica.

A situação é similar no caso do modelo log-linear usado para estimar as p_{im}'s. Nas tabelas de contingência não podem ser

consideradas as categorias para as quais duas ou mais variáveis binárias que representam a mesma variável categórica são não nulas, pois na realidade essas categorias não existem, (como ilustração revisar o exemplo 2.1.4). Felizmente, estes detalhes não afetam o algoritmo *iterative scaling*. Basta usar valores inicias nulos nas celas que representam as categorias inexistentes pois o uso do modelo log-linear garante que o elemento final nessas categorias também será zero (ver Haberman, 1972).

Uma vez que os termos apropriados são excluídos do modelo linear e que é dado como valor inicial zero às categorias apropriadas no algoritmo *iterative scaling*, o problema já está resolvido pois o procedimento de estimação e de classificação descritos antes podem ser usados sem inconveniências.

3.3 - FUNÇÃO LINEAR DISCRIMINANTE PARA MISTURAS

3.3.1 - Funcão Linear Discriminante de Fisher Adaptada para Misturas

Devido à falta de divulgação das técnicas para tratar problemas de discriminação com mistura de variáveis, este tipo de dados é frequentemente analisado usando métodos que originalmente foram desenvolvidos para variáveis contínuas. Destes o mais conhecido e difundido é o método de discriminação via Função Linear Discriminante (FLD) de Fisher, cuja adaptação para misturas foi formalizada por Krzanowski (1975). O atrativo da FLD de Fisher é a sua simplicidade como técnica e a facilidade de calculá-la, uma avaliação mais abrangente da FLD de Fisher pode ser encontrada no trabalho de Krzanowski (1977). Tudo isso somado ao fato de ter sido derivada sem fazer suposições distribucionais sugere o uso dela em problemas de classificação onde outras técnicas, provavelmente menos conhecidas,

poderiam dar melhores resultados.

Para começar o estudo da FLD de Fisher para o vetor misto **w** de q variáveis binárias e as p contínuas, definido na sub-seção 2.2.1 parte (a), suponha que a regra de alocação derivada seja:

$$\text{Alocar } \mathbf{w}' = (\mathbf{x}', \mathbf{y}') \text{ em } \pi_1 \text{ se } \mathbf{c}'\mathbf{w} + k \geq 0 \quad \text{e em } \pi_2, \text{ caso contrário.} \tag{3.3.1}$$

onde $\mathbf{c}' = (c_1, \ldots, c_q, c_{q+1}, \ldots, c_{p+q})$ é o vetor de coeficientes da FLD e k é a constante.

Para discriminar usando (3.3.1) é necessário conhecer as (p+q+1) constantes k, $c_1, \ldots, c_{p+q}$, para isso, recorde-se que a classificação de uma observação **w** com a FLD de Fisher é feita da seguinte maneira:

$$\text{Alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } (\nu_1 - \nu_2)' \, \Omega^{-1} \left(\mathbf{w} - \tfrac{1}{2}(\nu_1 + \nu_2)\right) \geq 0 \quad \text{e em } \pi_2 \text{ caso contrário.} \tag{3.3.2}$$

Em (3.3.2) ν_1 e ν_2 representam as médias de **w** em π_1 e π_2, respectivamente e Ω representa a matriz de dispersão que se supõe comum às duas populações, assim:

$$E(\mathbf{w}|\pi_i) = \nu_i \quad \text{e} \quad Cov(\mathbf{w}|\pi_i) = \Omega \qquad i = 1, 2$$

Igualando os termos correspondentes das equações (3.3.1) e (3.3.2), tem-se que:

$$\mathbf{c} = (\nu_1 - \nu_2)' \, \Omega^{-1} \quad \text{e} \quad k = -\tfrac{1}{2}(\nu_1 - \nu_2)' \, \Omega^{-1}(\nu_1 + \nu_2) \tag{3.3.3}$$

portanto, para discriminar usando (3.3.1) é necessário conhecer os

primeiros e os segundos momentos de **w** nas duas populações. Todos esses valores já foram calculados na parte (a) da sub-seção 2.2.1, logo a regra baseada na FLD de Fisher para mistura de variáveis está totalmente definida. Uma vez conhecida a regra de classificação, o passo seguinte é avaliar seu desempenho, para isso define-se o seguinte conjunto.

Seja D(m) o conjunto de variáveis binárias que toma valor 1 no m-ésimo arranjo possível do vetor $\mathbf{x} = (x_1, \ldots, x_q)'$; $m = 1, \ldots, 2^q$. Se $\xi(\mathbf{w})$ representa a FLD de Fisher para o vetor misto **w**, tem-se que de (3.1.1):

$$\xi(\mathbf{w}) = \mathbf{c}'\mathbf{w} + k,$$

onde **c** e k foram identificadas em (3.3.3), logo,

$$\xi(\mathbf{w}) = \sum_{j=1}^{q} c_j x_j + \sum_{j=1}^{p} c_{q+j} y_j + k$$

agora, dada uma observação $\mathbf{w} = (\mathbf{x}', \mathbf{y}')'$ tal que **x** assume o m-ésimo valor possível a_m, então denota-se $\xi(\mathbf{w} | \mathbf{x} = a_m) = \xi_m$, onde

$$\xi_m = \sum_{j \in D(m)} c_j + \sum_{j=1}^{p} c_{q+j} y_j + k, \qquad m = 1, \ldots, 2^q$$

Observe-se que as funções ξ_m representam 2^q hiperplanos paralelos. Logo, supondo que a ditribuição condicional de $\mathbf{y}' = (y_1, \ldots, y_p)$ dado que $\mathbf{x} = a_m$ seja normal p-variada, $N_p(\mu_i^{(m)}, \Sigma)$, em π_i (para i = 1, 2 e m = 1, ..., 2^q) então, a distribuição de ξ_m também será normal pois esta é uma combinação linear dos elementos de um vetor normal multivariado (ver Mardia *et al.*, 1979 p.41).

Calculando agora , os parâmetros da distribuição normal de ξ_m na população π_i tem-se que:

$$E(\xi_m) = E\left(\sum_{j \in D(m)} c_j + \sum_{j=1}^{p} c_{q+j}\, y_j + k \right)$$

$$= \sum_{j \in D(m)} c_j + \sum_{j=1}^{p} c_{q+j}\, \mu_{ij}^{(m)} + k \qquad (3.3.4)$$

$$Var(\xi_m) = Var\left(\sum_{j \in D(m)} c_j + c^{*\prime} y + k \right)$$

onde $c^{*} = (c_{q+1}, \ldots, c_{p+q})'$, então

$$Var(\xi_m) = c^{*\prime}\, Cov(y)\, c^{*} = c^{*\prime}\, \Gamma\, c^{*}, \qquad (3.3.5)$$

onde as componentes de Γ, γ_{ij} foram encontradas em (2.2.13) e (2.2.15), logo escrevemos:

$$\xi_m \sim N\left(\sum_{j \in D(m)} c_j + \sum_{j=1} c_{q+j}\, \mu_{ij}^{(m)} + k, \quad c^{*\prime}\, \Gamma\, c^{*} \right)$$

em π_i, $i = 1, 2$.

Finalmente, as probabilidades de classificar mal uma observação w cuja componente binária $x = a_m$ são:

$$P_m(2|1) = P(\xi_m < 0 \,|\, w \in \pi_1) = \Phi\left[\frac{k - c(m) - c^{*\prime}\, {}^{(m)}_{1}}{(c^{*\prime}\, \Gamma\, c^{*})^{1/2}} \right] \qquad (3.3.6)$$

$$P_m(1|2) = P(\xi_m \geq 0 \,|\, w \in \pi_2) = \Phi\left[\frac{c(m) + c^{*\prime}\, {}^{(m)}_{2} - k}{(c^{*\prime}\, \Gamma\, c^{*})^{1/2}} \right] \qquad (3.3.7)$$

onde, $c(m) = \sum_{j \in D(m)} c_j$, ${}^{(m)}_{i} = (\mu_{i1}^{(m)}, \ldots, \mu_{ip}^{(m)})'$ e Φ representa a função normal padrão acumulada.

De (3.3.6) e (3.3.7) segue-se que:

$$P(2|1) = \sum_{m=1}^{2^q} P_m(2|1)p_{1m} \quad \text{e} \quad P(1|2) = \sum_{m=1}^{2^q} P_m(1|2)p_{2m} \tag{3.3.8}$$

O valor dessas probabilidades permite avaliar a 'qualidade' com que seriam classificadas as observações de origem desconhecida usando a FLD de Fisher com vetores mistos.

Até agora toda análise foi feita para o caso de parâmetros conhecidos, porém na prática é mais comum ter que construir a regra de discriminação a partir de amostras de treinamento (*training samples*). Nesse caso, uma solução possível é substituir os parâmetros da FLD populacional pelos seus estimadores assim, a regra de decisão estimada seria:

$$\text{Alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } (\bar{\mathbf{w}}_1 - \bar{\mathbf{w}}_2)' \, S_w^{-1} \, (\mathbf{w} - \tfrac{1}{2}(\bar{\mathbf{w}}_1 + \bar{\mathbf{w}}_2)) \geq 0$$

$$\text{e em } \pi_2 \text{ caso contrário.} \tag{3.3.9}$$

onde,

$$\bar{\mathbf{w}}_i = \frac{1}{n} \sum_{j=1}^{n_i} w_{ij} \qquad i = 1, 2 \tag{3.3.10}$$

$$S_w = \frac{1}{n_1 + n_2 - 2} \sum_{i=1}^{2} \sum_{j=1}^{n_i} (w_{ij} - \bar{w}_i)(w_{ij} - \bar{w}_i)'$$

representam a média amostral das observações em π_i (i = 1, 2) e a matriz de covariância amostral ponderada, respectivamente.

Note-se que para obter as probabilidades de classificação erradas dadas de (3.3.6) a (3.3.8), também é necessário conhecer os parâmetros de **w**, portanto no caso destes serem desconhecidos o procedimento de

avaliação da regra de classificação deve ser estimativo, i.e., substituir os estimadores nas expressões correspondentes. Infelizmente, esta solução tende a ser otimista no sentido de subestimar as verdadeiras probabilidades de má classificação, quando o tamanho da amostra é pequeno (ver Mardia *et al.*, 1979 p. 321).

Existem outos métodos para estimar as probabilidades de classificação errada entre eles, o método de Lachenbruch (Lachenbruch, 1967), também chamado *leaving-one-out-method*, que está disponível, por exemplo no procedimento PROC DISCRIM do sistema de programas SAS. O problema de estimar as taxas de classificação errada tem ocupado muitos pesquisadores, algumas referências iniciais podem ser obtidas em Toussaint (1974).

3.3.2 - Funções Lineares Discriminantes Modificadas

Em 1982, Vlachonikolis e Marriott propuseram modificar a Função Linear Discriminante para melhorar o seu desempenho em problemas de classificação onde as variáveis envolvidas são binárias e contínuas. A idéia deles foi construir funções lineares discriminantes baseadas em alguns conjuntos de novas variáveis criadas a partir das variáveis observadas. O objetivo de introduzir as novas variáveis foi representar adequadamente na função de discriminação, o efeito devido às interações originadas pela presença de variáveis binárias. Esse interesse surge porque em trabalhos anteriores, os pesquisadores observaram que o desempenho da FLD de Fisher não é bom quando estruturas com acentuadas interações entre as variáveis binárias e/ou entre as variáveis binárias e os grupos, estão presentes nos dados (ver Krzanowski, 1975, 1977).

Nesta seção serão descritas em detalhe as três modificações propostas pelos autores.

a) Primeira Modificação de FLD

A primeira modificação é a mais simples e consiste construir uma nova FLD baseada no vetor misto $\mathbf{w}^* = (\mathbf{z}', \mathbf{y}')'$ definido na seção 2.2.1, parte (b). Nesse vetor as q variáveis binárias observadas originalmente $(x_1, \ldots, x_q)$, são transformadas em uma multinomial $\mathbf{z}' = (z_1, \ldots, z_r)$, $r = 2^q - 1$, usando a definição 2.1.7. A parte contínua do vetor original $\mathbf{y} = (y_1, \ldots, y_p)'$ é mantida sem alterações.

Para simplificar a notação, de aqui em diante o vetor $(\mathbf{z}', \mathbf{y}')'$ será denotado por $\mathbf{w}$ e não por $\mathbf{w}^*$.

A primeira função linear discriminante modificada pode ser escrita da seguinte maneira:

$$\xi(\mathbf{w}) = \alpha_1 z_1 + \ldots + \alpha_r z_r + \beta_1 y_1 + \ldots + \beta_p y_p + \beta_0 \qquad (3.3.11)$$

ou vetorialmente:

$$\xi(\mathbf{w}) = \beta_0 + \alpha'\mathbf{w} \qquad (3.3.12)$$

onde $\alpha = (\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_p)'$ e $\mathbf{w} = (z_1, \ldots, z_r, y_1, \ldots, y_p)'$.

A regra de discriminação baseada na função $\xi(\mathbf{w})$ dada em (3.3.12) é,

$$\text{Alocar } \mathbf{w} \text{ em } \pi_1 \text{ se } \xi(\mathbf{w}) = \beta_0 + \alpha'\mathbf{w} \geq 0 \qquad (3.3.13)$$
$$\text{e em } \pi_2 \text{ caso contrário.}$$

O número de coeficientes que deve ser estimado na função (3.3.11)

é $(r + p + 1) = (2^q + p)$ em lugar dos $(p + q + 1)$ que se teria se as variáveis contínuas e binárias fossem usadas diretamente. É interessante observar que se as médias e matrizes de variância-covarância de $\mathbf{w}$ em π_1 e π_2 são utilizadas para estimar esses coeficientes, o número de parâmetros a ser estimado cresce consideravelmente.

A função $\xi(\mathbf{w})$ em (3.3.11) tem uma forma particular para cada categoria definidas pelas variáveis binárias assim, na m-ésima categoria tem-se:

$$(\xi(\mathbf{w}) \mid z_m = 1) = \xi_m(\mathbf{w}) = \beta'\mathbf{y} + (\beta_0 + \alpha_m) \qquad (3.3.14)$$

$$m = 0, \ldots, r$$

onde $\beta' = (\beta_1, \ldots, \beta_p)$ e $\alpha_0 = 0$.

As funções $\xi_m(\mathbf{w})$ $(m = 0, \ldots, r)$, representam hiperplanos paralelos que separam os grupos, representados pelas variáveis contínuas, nas diferentes categorias. As 'posições' desses hiperplanos dependem do termo constante $(\beta_0 + \alpha_m)$, isto é, da categoria que foi observada.

Supondo que a distribuição condicional de $\mathbf{y}$ na m-ésima categoria de π_i tem média $\mu_i^{(m)}$ e matriz de dispersão Σ comum para $i = 1, 2$ e $m = 0, 1, \ldots, r$ então os momentos condicionais da função linear discriminante $\xi_m(\mathbf{w})$ definida em (3.3.14) são:

$$E(\xi_m(\mathbf{w})) = \beta'\mu_i^{(m)} + (\beta_0 + \alpha_m) \qquad (3.3.15)$$

e

$$\text{Var}\ (\xi_m(\mathbf{w})) = \beta'\Sigma\beta \qquad (3.3.16)$$

na população π_i $(i = 1, 2)$.

Sob hipótese de multinormalidade para $\mathbf{y}|z_m = 1$, $\xi_m(\mathbf{w})$ também tem distribuição normal pois é uma transformação linear de um vetor aleatório normal. Agora, as probabilidades de classificar uma observação da j-ésima população em π_i ($i \neq j$), estão dadas por

$$P(i|j) = \sum_{m=0}^{r} P_m(i|j)p_{jm}, \qquad i \neq j, \qquad i, j = 1, 2 \tag{3.3.17}$$

onde,

$$P_m(1|2) = P\left[\, \xi_m(\mathbf{w}) \geq 0 \,|\, \mathbf{w} \in \pi_2 \right]$$

$$= \Phi\left[\frac{\beta' \mu_2^{(m)} + (\beta_0 + \alpha_m)}{(\beta' \Sigma \beta)^{1/2}} \right] \tag{3.3.18}$$

e

$$P_m(2|1) = P\left[\, \xi_m(\mathbf{w}) \geq 0 \,|\, \mathbf{w} \in \pi_1 \right]$$

$$= \Phi\left[\frac{-\beta' \mu_1^{(m)} - (\beta_0 + \alpha_m)}{(\beta' \Sigma \beta)^{1/2}} \right] \tag{3.3.19}$$

para m = 0, ..., r. Como antes, Φ representa a distribuição acumulada de uma normal padrão e p_{jm}, a probabilidade de pertencer a m-ésima categoria da população π_j.

b) Segunda Modificação da FLD

A segunda modificação proposta por Vlachonikolis e Marriott (1982) foi construir uma função linear de discriminação nas componentes do vetor **v**, definido na parte (c) da seção 2.2.1. O vetor misto **v** tem como componentes as variáveis multinomiais z_1, ..., z_r (obtidas da transformação da binárias x_1, ..., x_q como na seção anterior), as p contínuas y_1, ..., y_p e os produtos da forma $y_j z_k$ (j =

1, ..., p; k = 1, ..., r) que representam as interações entre as variáveis contínuas e cada uma das categorias multinomiais geradas pelas variáveis binárias.

A função linear discriminante baseada no vetor **v** pode ser escrita como segue:

$$\psi(\mathbf{v}) = \alpha_1 z_1 + \ldots + \alpha_r z_r + \beta_1 y_1 + \ldots + \beta_p y_p + \gamma_{11} z_1 y_1 + \ldots + \gamma_{rp} z_r y_p + \beta_0$$

ou vetorialmente:

$$\psi(\mathbf{v}) = \mathbf{a}'\mathbf{v} + \beta_0 \qquad (3.3.20)$$

onde $\mathbf{a} = (\alpha_1, \ldots, \alpha_r, \beta_1, \ldots, \beta_p, \gamma_{11}, \ldots, \gamma_{rp})'$ e

$\mathbf{v} = (z_1, \ldots, z_r, y_1, \ldots, y_p, z_1y_1, \ldots, z_ry_p)'$.

Para que a função $\psi(\mathbf{v})$ dada em (3.3.20) esteja totalmente definida devem ser estimados $r + p + rp + 1 = 2^q(p + 1)$ coeficientes.

Agora, dada uma observação **v** na m-ésima categoria tem-se que a função $\psi(\mathbf{v})$ é:

$$(\psi(\mathbf{v}) \,|\, z_m = 1) = \psi_m(\mathbf{v}) = \alpha_m + \beta_1 y_1 + \ldots + \beta_p y_p + \gamma_{m1} y_1 + \ldots + \gamma_{mp} y_p + \beta_0$$

então,

$$\psi_m(\mathbf{v}) = (\beta_1 + \gamma_{m1})y_1 + (\beta_2 + \gamma_{m2})y_2 + \ldots + (\beta_p + \gamma_{mp})y_p + (\beta_0 + \alpha_m) \qquad (3.3.21)$$

ou vetorialmente:

$$\psi_m(\mathbf{v}) = \mathbf{b}_m'\mathbf{y} + (\beta_0 + \alpha_m) \qquad m = 0, 1, \ldots, r \qquad (3.3.22)$$

onde, $\mathbf{b}_m = (\beta_1 + \gamma_{m1}, \beta_2 + \gamma_{m2}, \ldots, \beta_p + \gamma_{mp})'$; $\alpha_0 = 0$ e $\gamma_{0j} = 0$ $\forall\ j = 1, \ldots, p$

Neste caso, as funções $\psi_m(\mathbf{v})$ dadas em (3.3.22) representam 2^q hiperplanos diferentes cuja 'posição' e 'inclinação' varia para cada categoria. Comparando $\xi_m(\mathbf{w})$ dada em (3.3.14) com $\psi_m(\mathbf{v})$ em (3.3.22), observa-se que o efeito de considerar os produtos da forma $z_j y_k$ na função discriminante é justamente o de modificar a 'inclinação' dos hiperplanos que separam as duas populações nas distintas categorias. Assim, tem-se que a inclusão dos termos z_l e $z_l y_j$ na função $\psi(\mathbf{v})$, em (3.3.20), assimila as interações entre variáveis binárias e as interações entres as variáveis binárias e contínuas melhorando seu desempenho em relação a FLD de Fisher adaptada para misturas, estudada na seção 3.3.1.

É interessante notar que esta modificação é similar ao enfoque baseado no modelo de posição (seção 3.2), no sentido que produz uma função linear discriminante diferente para cada categoria . Porém, a modificação é mais geral pois não é necessário supor igualdade de matrizes de variância-covariância dentro das categorias.

Vlachonikolis e Marriott (1982) afirmam:

" ... o caso de igualdade de matrizes de covariância pode ser construído neste modelo, estimando todas as matrizes de dispersão das $z_j\mathbf{y}$'s, $j = 1, \ldots, r$, pela matriz de covariância de $\mathbf{y}$ dentro dos grupos."

entretanto, foi encontrado que a matriz de variância-covariância do vetor:

$$z_j\mathbf{y} = \begin{cases} \mathbf{y} & \text{se } z_j = 1 \\ 0 & \text{se } z_j = 0 \end{cases}$$

não representa a dispersão das variáveis contínuas dentro da j-ésima categoria senão função linear dela. Isto pode ser observado na expressão (2.2.44) da seção 2.2.1 que apresenta-se novamente a seguir, considerando agora as populações. Em π_i tem-se:

$$\text{Cov}\,(z_j y_k,\ z_j y_l) = \sigma_{kl}^{(j)} p_{ij} + \mu_{ik}^{(j)} \mu_{il}^{(j)}\ p_{ij}(1 - p_{ij})$$

$$j = 1, \ldots, r; \quad k, l = 1, \ldots, p; \quad i = 1, 2 \qquad (3.3.23)$$

onde

$$\sigma_{kl}^{(j)} = \text{Cov}\,(y_k,\ y_l | z_j = 1) \qquad \text{em } \pi_1 \text{ ou } \pi_2$$

$$\mu_{it}^{(j)} = E(y_t | z_j = 1,\ \pi_i), \qquad t = k, l \quad \text{e}$$

$$p_{ij} = P(z_j = 1 | \pi_i)$$

Portanto, o único efeito de supor que as matrizes de dispersão de $\mathbf{y}$ dentro das categorias são iguais é que a covariância anterior pode ser escrita como segue:

$$\text{Cov}\,(z_j y_k,\ z_j y_l) = \sigma_{kl} p_{ij} - \mu_{ik}^{(j)} \mu_{il}^{(j)}\ p_{ij}(1 - p_{ij})$$

$$j = 1, \ldots, r; \quad k, l = 1, \ldots, p; \quad i = 1, 2 \qquad (3.3.24)$$

No caso amostral, o estimador da matriz de dispersão de $z_j \mathbf{y}$ em π_i, (conforme foi visto em (2.2.70) da seção 2.2.2) é:

$$S^2_{z_j \mathbf{y}} = \frac{1}{n_i}\, \mathbf{y}_i^{(j)'} \left(I_{n_{ij}} - \frac{1}{n_i} 1_{n_{ij}} 1'_{n_{ij}} \right) \mathbf{y}_i^{(j)}$$

que pode ser escrito como:

$$S^2_{z_j y} = \frac{n_{ij}}{n_i} S_i^{2(j)} + \frac{n_{ij}}{n_i}\left(1 - \frac{n_{ij}}{n_i}\right) \bar{y}_i^{(j)} \bar{y}_i^{(j)'} \qquad (3.3.25)$$

onde:

$Y_i^{(j)}$: matriz de observação de y que pertence a j-ésima categoria de π_i

$\bar{y}_i^{(j)}$: média amostral de y na categoria j de π_i

n_{ij} : número de observação na i-ésima categoria de π_i

$S_i^{2(j)}$: estimador da matriz de dispersão de y na categoria j de π_i

$j = 1, \ldots, r \qquad i = 1, 2$

Portanto, uma proposta alternativa à dos autores para tratar o caso de igualdade de matrizes de covariância dentro das categorias seria substituir em (3.3.25) o termo $S_i^{2(j)}$ pelo estimador usual da matriz de covariância dentro, **W**, dado por:

$$\mathbf{W} = \frac{1}{n} \sum_{i=1}^{2} \sum_{j=1}^{2^q} n_{ij} S_i^{2(j)} \qquad (3.3.26)$$

Assim, a covariância amostral de $z_j y$ em π_i, estaria dada por:

$$S^2_{z_j y} = \frac{n_{ij}}{n_i} \left(\mathbf{W} - \left(1 - \frac{n_{ij}}{n_i}\right) \bar{y}_i^{(j)} \bar{y}_i^{(j)'} \right) \qquad (3.3.27)$$

$j = 1, \ldots, r \quad e \quad i = 1, 2$

Continuando com o problema de discriminação tem-se que a regra de classificação baseada na segunda função linear modificada $\psi(\mathbf{v})$ é:

dado que $z_m = 1$,

alocar a observação em π_1 se $\psi_m(\mathbf{v}) = \mathbf{b}'_m \mathbf{y} + (\beta_0 + \alpha_m) \geq 0$

e em π_2 caso contrário, $\qquad m = 0, 1, \ldots, r$

onde,

$\psi_m(\mathbf{v})$ foi definida em (3.3.22) e

$$z_0 = \begin{cases} 1 \text{ se} & z_j = 0 \quad \forall\ j = 1, \ldots, r \\ 0 \text{ se} & \exists\ z_j = 1 \quad j = 1, \ldots, r \end{cases}$$

foi definido em (2.2.17) da seção 2.2.1.

Agora, supondo que a distribuição condicional de $\mathbf{y}$ na m-ésima categoria de π_1 é normal com média $\mu_1^{(m)}$ e matriz de variância-covariância $\Sigma^{(m)}$, tem-se que a função $\psi_m(\mathbf{v})$ também é normal com parâmetros:

$$E(\psi_m(\mathbf{v})) = \mathbf{b}_m'\mu_1^{(m)} + (\beta_0 + \alpha_m) \tag{3.3.28}$$

e

$$\mathrm{Var}(\psi_m(\mathbf{v})) = \mathbf{b}_m'\Sigma^{(m)}\mathbf{b}_m. \tag{3.3.29}$$

As probabilidades de classificar mal uma observação da m-ésima categoria, sob hipótese de normalidade são:

$$P_m(1|2) = \Phi\left[\frac{\mathbf{b}_m'\mu_2^{(m)} + (\beta_0 + \alpha_m)}{(\mathbf{b}_m'\Sigma^{(m)}\mathbf{b}_m)^{1/2}}\right] \tag{3.3.30}$$

$$P_m(2|1) = \Phi\left[-\frac{\mathbf{b}_m'\mu_1^{(m)} + (\beta_0 + \alpha_m)}{(\mathbf{b}_m'\Sigma^{(m)}\mathbf{b}_m)^{1/2}}\right] \tag{3.3.31}$$

$$m = 0, 1, \ldots, r.$$

Finalmente, as probabilidades totais de má classificação são:

$$P(i|j) = \sum_{m=0}^{r} P_m(i|j)\, p_{jm}, \qquad j \neq i \quad j, i = 1, 2$$

c) Terceira Modificação da FLD

Os dois modelos modificados propostos anteriormente requerem da estimação de um número muito grande de parâmetros a menos que o número de variáveis binárias seja pequeno. Uma maneira de enfrentar este problema poderia ser utilizando aproximações similares às propostas em Krzanowski (1975) para o modelo de posição. Porém, existe outra forma mais compatível com este enfoque que também foi proposta em Vlachonikolis e Marriott (1982).

A idéia dos autores foi representar as interações devidas às variáveis binárias usando diretamente estas variáveis e não a transformação multinomial que era usada nas outras modificações. Neste caso, a função linear discriminante é construída considerando todos os termos que representam efeitos principais e interações de primeira e segunda ordem das variáveis $x_1, \ldots, x_q$ e $y_1, \ldots, y_p$. É claro que também poderiam ser incluídas as interações de ordem maior, mas isto não contribuiria em nada a resolver o problema de excessivos parâmetros.

A função linear discriminante modificada para o modelo reduzido é:

$$\phi(\mathbf{w}) = \sum_{i=1}^{q} \alpha_i x_i + \sum_{i=1}^{p} \beta_i y_i + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \gamma_{ij} x_i x_j + \sum_{i=1}^{q} \sum_{j=1}^{p} \varepsilon_{ij} x_i y_j + \sum_{i=1}^{q-1} \sum_{j=i+1}^{q} \sum_{k=1}^{p} \delta_{ijk} x_i x_j y_k + \beta_0 . \qquad (3.3.32)$$

Em $\phi(\mathbf{w})$, (3.3.32), devem ser estimados $q + \binom{q}{2} + p + qp + p\binom{q}{2} + 1 = (p + 1)(q^2/2 + q/2 + 1)$ coeficientes. Quando q cresce, este número é consideravelmente menor que o número de coeficientes que deviam ser estimados nas modificações vistas nas partes (a) e (b) desta seção. A tabela 3.3.1 mostra este fato claramente.

Tabela 3.3.1: Comparação de número de coeficientes da FLD simples e as três modificações estudadas nessa seção (p fixo).

q	FLD	$FLDM_1$	$FLDM_2$	$FLDM_3$
1	p + 2	p + 2	2(p + 1)	2(p + 1)
2	p + 3	p + 4	4(p + 1)	4(p + 1)
3	p + 4	p + 8	8(p + 1)	7(p + 1)
5	p + 6	p + 32	32(p + 1)	16(p + 1)
10	p + 11	p + 1024	1024(p + 1)	56(p + 1)
15	p + 16	p + 32768	32768(p + 1)	121(p + 1)

Todas as modificações apresentadas têm a vantagem de representar no modelo as interações devidas às variáveis binárias. Além disto, permitem usar diretamente os métodos *stepwise* de seleção de variáveis implementados nos pacotes computacionais, a fim de obter informação sobre o poder de discriminação de cada variável. Por exemplo, a seleção de z_i e $z_i y_k$ pode ser interpretada como evidência de que o efeito principal do padrão de $\mathbf{x}$ representado por z_i é significativo e que a interação de z_i com y_k também o é. Interpretações similares se aplicam aos modelos reduzidos.

3.4 - DISCRIMINAÇÃO LOGÍSTICA

Os métodos de discriminação são usados principalmente com dois objetivos, primeiro para resumir e descrever diferenças entre grupos e segundo para alocar novas observações nestes grupos. O interesse fundamental da discriminação logística é o segundo, isto é, alocar um indivíduo com vetor de observações $\mathbf{w}$ em uma das duas populações π_1 ou π_2.

A discriminação logística pode ser descrita como um método

parcialmente distribucional, pois sua suposição fundamental para resolver problemas de classificação, é a linearidade do logarítmo da razão das verosimilhanças. Algebricamente essa condição pode ser escrita como segue:

$$\ln\left[\frac{P(\mathbf{w}|\pi_1)}{P(\mathbf{w}|\pi_2)}\right] = \alpha + \beta'\mathbf{w} \qquad (3.4.1)$$

onde $\mathbf{w} = (w_1, \ldots, w_r)'$, $\beta = (\beta_1, \beta_2, \ldots, \beta_r)'$ e $P(\mathbf{w}|\pi_i)$, $i = 1, 2$, representa a verossimilhança de $\mathbf{w}$ em π_i.

Esta suposição é importante e útil pois quando usada junto com o teorema de Bayes permite escrever as probabilidades posteriores de forma simples. Assim, pelo teorema de Bayes, tem-se

$$P(\pi_1|\mathbf{w}) = \frac{P(\mathbf{w}|\pi_1)\,P(\pi_1)}{P(\mathbf{w}|\pi_1)\,P(\pi_1) + P(\mathbf{w}|\pi_2)P(\pi_2)}$$

$$= \left[1 + \frac{P(\mathbf{w}|\pi_2)\,P(\pi_2)}{P(\mathbf{w}|\pi_1)\,P(\pi_1)}\right]^{-1}$$

logo substituindo (3.4.1) na expressão anterior:

$$P(\pi_1|\mathbf{w}) = \frac{1}{1 + [k \exp(\alpha + \beta'\mathbf{w})]^{-1}},$$

então,

$$P(\pi_1|\mathbf{w}) = \frac{k \exp(\alpha + \beta'\mathbf{w})}{1 + k \exp(\alpha + \beta'\mathbf{w})} \qquad (3.4.2)$$

onde $k = \frac{P(\pi_1)}{P(\pi_2)}$, representa o *odds* de pertencer à população π_1.

Agora,

$$P(\pi_2|w) = 1 - P(\pi_1|w) = \frac{1}{1 + k \exp(\alpha + \beta'w)} \qquad (3.4.3)$$

Logo, assumindo que α, β e k já tenham sido estimados, a regra ótima, que consiste em alocar uma nova observação à população com maior probabilidade à posteriori, será fácil de usar e calcular pois dependerá unicamente da função linear $(\alpha + \beta'w + \ln k)$.[1]

A estimação dos parâmetros α e é feita usando o método de máxima verossimilhança. A vantagem de usá-lo é que o método não varia para os diferentes tipos de variáveis, isto significa que embora as variáveis sejam contínuas, discretas ou mistura dos dois tipos o procedimento de estimação será o mesmo.

É interessante observar que até agora, a única restrição distribucional feita para o enfoque logístico foi colocada em (3.4.1). Isso garante a utilidade do método, pois existem muitas famílias distribucionais que satisfazem essa restrição (para maiores detalhes ver Anderson, 1972 e 1975), entre elas têm-se:

i) Distribuições normais multivariadas com matrizes de covariância iguais.

ii) Distribuições discretas multivariadas seguindo o modelo log-linear com interações iguais nos dois grupos.

---

(1) Note-se que $k \exp(\alpha + \beta'W) = \exp(\ln k + \alpha + \beta'W)$

iii) Distribuições discretas multivariadas independentes.

iv) Distibuições conjuntas de variáveis contínuas e categóricas seguindo i) e ii), não necessariamente independentes.

v) Algumas versões truncadas de casos anteriores.

vi) Versões dos casos anteriores onde **w**, é substituído por alguma função de **w** em (3.4.1).

A forma de função de verossimilhança, que deverá ser maximizada para estimar os parâmetros, depende do tipo de amostragem usada para obter as amostras de treinamento. Serão estudados três delineamentos, que são comuns na prática e que produzem a mesma função de verossimilhança, este fato simplifica muito o processo de maximização.

Seguindo a terminologia usada por Anderson (1982), tem-se:

- Amostragem condicional em **w**

A amostragem condicional consiste em extrair uma ou mais amostras para cada valor fixo de **w**; cada uma dessas amostras pode assumir dois valores π_1 ou π_2, dependendo da população à qual pertença a observação. Este procedimento representa uma amostragem da distribuição condicional de $\pi|\mathbf{w}$ com verossimilhança $P(\pi|\mathbf{w})$.

Supondo que uma amostra de tamanho n seja escolhida, e que dado **w**, há $n_i(\mathbf{w})$ pontos da população π_i (i = 1, 2) que assumem esse valor então, sob a amostragem condicional em **w**, a verossimilhança é,

$$L_c = \prod_{\mathbf{w}} \{ P(\pi_1|\mathbf{w})\}^{n_1(\mathbf{w})} \{ P(\pi_2|\mathbf{w})\}^{n_2(\mathbf{w})} \qquad (3.4.4)$$

Aqui, o valor $n(\mathbf{w}) = n_1(\mathbf{w}) + n_2(\mathbf{w})$ é fixo para todo **w**. Logo, fazendo $\beta_0 = \alpha + \ln k$ em (3.2.2), tem-se que:

$$P(\pi_1|\mathbf{w}) = \frac{\exp(\beta_o + \beta'\mathbf{w})}{1 + \exp(\beta_o + \beta'\mathbf{w})} = p_1(\mathbf{w})$$

e em (3.4.3):

$$P(\pi_2|\mathbf{w}) = \frac{1}{1 + \exp(\beta_o + \beta'\mathbf{w})} = p_2(\mathbf{w})$$

A verossimilhança, L_c, pode ser escrita agora como:

$$L_c = \prod_{\mathbf{w}} \{ p_1(\mathbf{w})\}^{n_1(\mathbf{w})} \{ p_2(\mathbf{w})\}^{n_2(\mathbf{w})} \qquad (3.4.5)$$

Substituindo em (3.4.5), $p_1(\mathbf{w})$ e $p_2(\mathbf{w})$ por seus respectivos valores é fácil ver que L_c é função de β_0 e β_j (j = 1, ..., p). Os estimadores de máxima verossimilhança destes parâmetros serão derivados usando procedimentos de otimização iterativa, como o Método de Newton-Raphson ou os métodos quase-Newton. Algo importante a ser observado é que β_0 é estimável, mas $\alpha = \beta_0 - \ln k$ não, a menos que $P(\pi_1)$ seja conhecido ou possa ser estimado separadamente.

- Amostragem mista

Neste tipo de amostragem, as observações são extraídas da distribuição conjunta $(\pi, \mathbf{w})$ com verossimilhança $L(\pi, \mathbf{w})$, onde π assume valores π_1 ou π_2. A proporção de elementos de π_1 na amostra, é uma estimativa da probabilidade prévia de pertencer a π_1, $P(\pi_1)$, e a partir dela $P(\pi_2) = 1 - P(\pi_1)$ também é estimável.

Sob amostragem mista, a verossimilhança é dada por:

$$L_m = \prod_{\mathbf{w}} \{ P(\mathbf{w}|\pi_1)\}^{n_1(\mathbf{w})} \{ P(\mathbf{w}|\pi_2)\}^{n_2(\mathbf{w})} \qquad (3.4.6)$$

Embora L_m seja diferente de L_c, é fácil demonstrar que neste caso, β_0 e também podem ser estimados maximizando a verossimilhança L_c dada em (3.4.5). Observe-se que:

$$P(\mathbf{w}|\pi_s) = P(\pi_s|\mathbf{w})\ P(\mathbf{w}) \qquad s = 1, 2,$$

então,

$$L_m = \prod_{\mathbf{w}} [P(\pi_1|\mathbf{w})P(\mathbf{w})]^{n_1(\mathbf{w})}[P(\pi_2|\mathbf{w})\ P(\mathbf{w})]^{n_2(\mathbf{w})}$$

$$= \prod_{\mathbf{w}} [P(\pi_1|\mathbf{w})]^{n_1(\mathbf{w})}[P(\pi_2|\mathbf{w})]^{n_2(\mathbf{w})} \prod_{\mathbf{w}} (P(\mathbf{w}))^{n(\mathbf{w})}$$

substituindo (3.4.4) na expressão anterior, tem-se que:

$$L_m = L_c \prod_{\mathbf{w}} (P(\mathbf{w}))^{n(\mathbf{w})} \qquad (3.4.7)$$

Refletindo sobre a expressão anterior pode-se concluir que somente L_c contém informação sobre os parâmetros, pois a única suposição feita neste enfoque é sobre a forma funcional da razão das verossimilhanças, isto leva a pensar que a distribuição marginal de $\mathbf{w}$ não contribui à estimação de β_0 e Anderson (1982), afirma que se $\prod_{\mathbf{w}}(P(\mathbf{w}))^{n(\mathbf{w})}$ tivesse alguma informação adicional sobre a forma de $P(\mathbf{w}|\pi_i)$, a informação extra sobre os parâmetros obtida dela seria pequena comparada com a contida em L_c. Assim, tem-se mostrado que para a amostragem mista os estimadores de máxima verossimilhança também são obtidos maximizando L_c em (3.4.5). É interessante notar que neste caso os valores $P(\pi_i)$ i = 1, 2 e α são estimáveis.

- Amostragem separada

Este tipo de amostragem é a mais comum nos problemas de discriminação; o método consiste em extrair amostras de cada população

por separado, ou em outras palavras, em realizar uma amostragem das distribuições condicionais, $(\mathbf{w}|\pi_1)$ e $(\mathbf{w}|\pi_2)$.

Sob amostragem separada, a verossimilhança das observações é:

$$L_s = \prod_{s=1}^{2} \prod_{w} \{ P(\mathbf{w}|\pi_s) \}^{n_s(w)} \qquad (3.4.8)$$

Anderson (1972), provou que no caso de variáveis discretas, maximizar L_s em (3.4.8), é equivalente a maximizar L_c em (3.4.5) porém, quando alguma ou todas as variáveis são contínuas não é possível provar esta equivalência. No mesmo artigo, o autor propõe que em presença de variáveis contínuas, ainda seja usado o tratamento anterior, isto é, estimar os parâmetros maximizando L_c. A justificativa mais simples para isso baseia-se na possibilidade de subdividir cada variável contínua para fazê-la discreta (embora alguma informação seja perdida pela discretização).

Posteriormente, Anderson e Blair (1982) mostraram que para variáveis contínuas, os estimadores obtidos a partir de L_c tecnicamente já não são de máxima verossimilhança. Isto os motivou para sugerir um método alternativo chamado máxima verossimilança penalizada (*penalized maximum likelihood*).

É importante recordar que os problemas com variáveis contínuas só aparecem na amostragem separada. Para amostragem condicional ou mista a estimação de β_0 e a partir de L_c é inquestionável sem importar a natureza das variáveis.

Resumindo, tem-se que na discriminação logística, os parâmetros necessários para calcular as probabilidades à posteriori, em muitos casos podem ser estimados usando o mesmo processo de máxima verossimilhança para dados discretos e/ou contínuos, para diferentes famílias de distribuições e para distintos planos de amostragem; em

todas essas situações a função a maximizar é:

$$L_c = \prod_w \{P(\pi_1 | w)\}^{n_1(w)} \{ P(\pi_2 | w)\}^{n_2(w)}$$

Agora que já é conhecida a função que deve ser maximizada, a etapa seguinte é escolher um procedimento adequado. Day e Kerridge (1967) e Anderson (1972) propuseram originalmente usar o procedimento de Newton-Raphson para maximizar L_c, porém o método pode não convergir (ver Jones, 1975). Anderson (1982), notou que os métodos quase-Newton são mais apropriados pois combinam a velocidade de convergência para valores próximos do ótimo do método de Newton com algumas propriedades do *Steepest Descent Method* para valores iniciais pobres (ver Gill e Murray, 1972). Outras vantagens dos métodos quase-Newton é que requerem somente as primeiras derivadas em cada iteração, e além disso que produzem uma estimativa da matriz de segundas derivadas no ponto máximo. Isto é conveniente na estatística, pois assim, é possível conhecer a matriz de informação sem calculá-la explicitamente. Anderson (1982) afirma também que para um número adequado de iterações (que não seja menor que o número de variáveis), o erro introduzido pelo método quase-Newton nas variâncias assintóticas estimadas é de ordem de 5%, que é um valor aceitável na maioria dos casos.

Agora, aplicando logarítmo natural à função L_c tem-se:

$$\ln(L_c) = \sum_{l=1}^{2} \sum_w n_l(w) \ln (P(\pi_l | w))$$

$$= \sum_w \{ n_1(w)(\beta_o + \beta' w) + n_1(w) \ln P(\pi_2 | w) + n_2(w) \ln P(\pi_2 | w)\}$$

$$= \sum_w \{ n_1(w)(\beta_o + \beta' w) + n(w) \ln P(\pi_2 | w)\}$$

onde $n(w) = n_1(w) + n_2(w)$. Logo as equações de máxima verossimilhança

são:

$$\frac{\partial \ln(L_c)}{\partial \beta_k} = \sum_w \left(n_1(w) w_k + n(w)(-1) \frac{\exp(\beta_0 + \beta' w)\, w_k}{1 + \exp(\beta_0 + \beta' w)} \right)$$

$$= \sum_w \{ n_1(w) - n(w) P(\pi_1 | w) \}\, w_k = 0 \qquad (3.4.9)$$

$$k = 0, 1, \ldots, r$$

onde $w = (w_1, w_2, \ldots, w_r)$ e $w_0 = 1$

$$\frac{\partial^2 \ln(L_c)}{\partial \beta_l \partial \beta_k} = - \sum_w \frac{n(w)\, w_l w_k \{ \exp(\beta_0+\beta' w)\} \{ 1+\exp(\beta_0+\beta' w) - \exp(\beta_0+\beta' w)\}}{[\, 1 + \exp(\beta_0+\beta' w)\,]^2}$$

$$= - \sum_w n(w)\, w_l w_k \frac{\exp(\beta_0+\beta' w)}{1 + \exp(\beta_0+\beta' w)} \cdot \frac{1}{1 + \exp(\beta_0+\beta' w)}$$

$$\therefore \quad - \frac{\partial^2 \ln(L_c)}{\partial \beta_l \beta_k} = \sum_w n(w) P(\pi_1 | w) P(\pi_2 | w)\, w_l w_k \qquad (3.4.10)$$

$$k, l = 0, \ldots, r$$

A única informação que está faltando para começar os processos de otimização iterativa Newton-Raphson ou quase-Newton, é sobre os valores iniciais. Cox (1966) sugeriu aproximações lineares de $P(\pi_i | w)$, $i = 1, 2$, e obter estimativas iniciais de β_0 e por mínimos quadrados ponderados. Porém em 1972, Anderson sugeriu começar usando zero como valor inicial para os $p + 1$ parâmetros logísticos e algum tempo depois afirmou que isto tem funcionado bem na prática e pode ser recomendado com confiança (ver Anderson, 1982). Por outro lado, Albert (1978)

mostrou que a função de verossimilhança L_c tem um único máximo, alcançado para β finito, exceto sob duas circunstâncias especiais facilmente reconhecíveis:

i) Separação completa, na qual todos os pontos de π_1 caem num lado de um hiperplano e os pontos de π_2 do outro lado. Day e Kerridge (1967) e Anderson (1972), tratam este caso em detalhe.

ii) Proporções marginais nulas, quando as variáveis são discretas. Anderson (1974), analisa este caso e encontra estimadores apropriados.

Por último, é importante notar que o processo de maximização iterativa de L_c é aplicável aos três tipos de amostragem quando usado para encontrar estimadores de máxima verossimilhança de β_0 e β. Para amostragem mista e condicional, a variância assintótica dos estimadores está dada pelos elementos da diagonal de $\mathbf{A}^{-1}$, onde

$$\mathbf{A} = \left(\frac{\partial^2 \ln(L_c)}{\partial \beta_e \beta_k} \right)$$

(ver 3.4.10), é a matriz de informação onde os parâmetros são substituídos pelos valores que maximizam a função de verossimilhança. No caso da amostragem separada são necessárias restrições adicionais (ver Anderson (1982) ou Seber (1984)).

3.5 - OUTROS MÉTODOS

3.5.1 - Discriminação Kernel

Alguns dos métodos discutidos até agora dependem, de alguma forma, de certas suposições sobre a distribuição do vetor aleatório

misto $\mathbf{w}$ nas populações π_1 e π_2. Se algumas destas suposições fossem relaxadas, seria possível usar métodos não paramétricos para estimar as densidades e depois derivar a regra de classificação. Neste contexto, vários autores como Habbema *et al.* (1974) e Aitchinson e Aitken (1976), tem usado os estimadores Kernel de densidades multivariadas (ver também Cacoullos, 1966; Murthy, 1966 e Breiman *et al.*, 1977).

Embora a teoria sobre os métodos de estimação Kernel esteja bastante desenvolvida, fazer uma revisão detalhada dela não está dentro do alcance deste trabalho, por tanto, o método de discriminação Kernel será descrito brevemente e algumas referências bibliográficas iniciais serão dadas.

Para começar, apresenta-se um estimador Kernel da função densidade $f(\mathbf{y})$, de alguma distribuição desconhecida, baseado no conjunto $\mathbf{D}$ de n observações independentes desta distribuição, $\mathbf{x}_1, \ldots, \mathbf{x}_n$.

Seja $K(\mathbf{y}|\mathbf{x}, \lambda)$, que denota uma classe de funções de densidade de $\mathbf{y}$ com moda em $\mathbf{x}$ e parâmetro de alisamento (ou janela) λ, cujo valor depende da população π_1. Uma versão comum do método Kernel é usar como estimador de $f(\mathbf{y})$, uma mistura de densidades Kernel tal como:

$$\hat{f}(\mathbf{y}) = P(\mathbf{y}|\mathbf{D}, \lambda) = \frac{1}{n} \sum_{j=1}^{n} K(\mathbf{y}|\mathbf{x}_j, \lambda) \qquad (3.5.1)$$

onde $\mathbf{y} = (y_1, y_2, \ldots, y_d)$ e $\mathbf{x}_j = (x_{j1}, \ldots, x_{jd})$, $j = 1, \ldots, n$.

Continuando, será revisada brevemente a discriminação Kernel para variáveis contínuas e logo para variáveis discretas, pois o tratamento particular para variáveis mistas surge como extensão destes dois métodos.

a) Dados Contínuos:

Quando as variáveis explanatórias são contínuas, a função de densidade normal multivariada é frequentemente usada como função Kernel em (3.5.1) devido as suas propriedades de escala e de unimodalidade. O Kernel multinormal é o seguinte:

$$K(\mathbf{y}|\mathbf{x}, \lambda) = \frac{1}{(2\pi\lambda)^{d/2}} |S|^{-1/2} \exp\left(\frac{-1}{2\lambda^2} (\mathbf{y} - \mathbf{x})' S^{-1} (\mathbf{y} - \mathbf{x}) \right) \qquad (3.5.2)$$

onde $S = \frac{1}{n-1} \sum_{j=1}^{n} (\mathbf{x}_j - \bar{\mathbf{x}})(\mathbf{x}_j - \bar{\mathbf{x}})'$ e $\bar{\mathbf{x}} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i$

Habbema *et al.* (1974) propuseram usar a matriz diagonal de elementos de S, diag(S), e não S diretamente, nesse caso o Kernel multinormal seria:

$$K(\mathbf{x}|\mathbf{y}, \lambda) = \frac{1}{(2\pi\lambda^2)^{d/2}} \left(\prod_{k=1}^{d} S_k^2 \right)^{-1} \exp\left(-\frac{1}{2} \sum_{k=1}^{d} \left(\frac{y_k - x_k}{\lambda S_k} \right)^2 \right) \qquad (3.5.3)$$

onde

$$(n-1)\, S_k^2 = \sum_{j=1}^{n} (x_{jk} - \bar{x}_k), \quad k = 1, \ldots, d;$$

$$\mathbf{y} = (y_1, \ldots, y_d), \mathbf{x} = (x_1, \ldots, x_d);$$

e $\quad \mathbf{x}_j = (x_{j1}, x_{j2}, \ldots, x_{jd}), \; j = 1, \ldots, n.$

Agora, se algum estimador $\hat{\lambda}$, da janela λ for conhecido, $f(\mathbf{y})$ pode ser estimada por:

$$\hat{f}(\mathbf{y}) = P(\mathbf{y}|D, \hat{\lambda}) = \frac{1}{n} \sum_{j=1}^{n} K(\mathbf{y}|\mathbf{x}_j, \hat{\lambda})$$

Murthy (1966), mostrou que este estimador é consistente para todos os pontos de continuidade de $\mathbf{y}$ se $\hat{\lambda} \to 0$ tão lentamente quanto $n \to \infty$.

Quando $\mathbf{y}$ é alguma das obserações de D, digamos $\mathbf{y} = \mathbf{x}_r$, então $f(\mathbf{x}_r)$ estima-se por:

$$\hat{f}^{(r)}(\mathbf{x}_r) = \frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq r}}^{n} K(\mathbf{x} | \mathbf{x}_r, \lambda_j) \qquad (3.5.4)$$

A estimação de uma janela λ adequada é um dos problemas mais importantes da estimação Kernel; infelizmente, o método de máxima verossimilhança produz um estimador nulo quando a função de verossimilhança $\prod_{j=1}^{n} P(\mathbf{x}_j | D, \lambda)$ é considerada. Habbema *et al.*, (1974) estudaram este problema e sugeriram usar um método de máxima verossimilhança modificado conhecido como método verossimilhança Jacknife (*Jacknife Likelihood Method*) ou *leaving one out modified maximum likelihood*, que consiste em maximizar, em relação a λ, a função $h(\lambda)$ dada por:

$$h(\lambda) = \prod_{r=1}^{n} \left[\frac{1}{n-1} \sum_{\substack{j=1 \\ j \neq r}}^{n} \tilde{K}(\tilde{\mathbf{x}}_r | \tilde{\mathbf{x}}_j, \lambda) \right]$$

onde

$$\tilde{\mathbf{x}}_k = \left(\frac{x_{k1}}{s_1}, \ldots, \frac{x_{kd}}{s_d} \right) \qquad k = 1, \ldots, n$$

representa os dados padronizados e

$$\tilde{K}\,(\tilde{x}_r|\tilde{x}_j,\ \lambda) = \frac{1}{(2\pi\lambda^2)^{d/2}}\ \exp\left[-\frac{1}{2\lambda^2}\,(\tilde{x}_r - \tilde{x}_j)'(\tilde{x}_r - \tilde{x}_j)\right]$$

$$\forall\ j \neq r \quad e\ j, r = 1, \ldots, n$$

No problema de discriminação, o método anterior pode ser usado para estimar $f_1(w)$ e $f_2(w)$, as densidades de w em π_1 e π_2 respectivamente, logo, a regra de classificação seria:

$$\text{Alocar } w \text{ em } \pi_1 \text{ se } \frac{\hat{f}_1(w)}{\hat{f}_2(w)} > \delta \qquad (3.5.5)$$

$$\text{e em } \pi_2 \text{ caso contrário.}$$

Em (3.5.5), δ é uma constante que depende dos custos de classificação errada e das probabilidades prévias.

Habbema *et al.* (1974) usaram um procedimento *forward* para seleção de variáveis, o qual, segundo Seber (1984) não é recomendável para problemas de grande porte pois consume muito tempo computacional.

É importante observar que, embora tenha sido usado um Kernel multivariado, não foi feita suposição nenhuma sobre a densidade a ser estimada, portanto, o método é estritamente não paramétrico e além disso qualquer outra densidade unimodal pode ser usada como Kernel.

b) Dados Binários

Aitchinson e Aitken (1976), sugeriram usar a seguinte função Kernel para dados binários multivariados:

$$K(y|x,\ \lambda) = \lambda^{d-D(y,x)}\ (1-\lambda)^{D(y,x)} \qquad (3.5.6)$$

onde $\frac{1}{2} \leq \lambda \leq 1$ e $D(y,x) = \|y - x\|^2$

Note-se que, quando as variáveis são binárias, $D(\mathbf{x}, \mathbf{y})$ representa simplesmente, o número de diferenças entre os elementos correspondentes de $\mathbf{y}$ e $\mathbf{x}$. A expressão (3.5.6) pode ser escrita de maneira mais simples se a tranformação $w_k = |y_k - x_k|$, $k = 1, \ldots, d$, de $\mathbf{y}$ em $\mathbf{w}$ para um $\mathbf{x}$ fixo é usada. Substituindo $\mathbf{w}$ em (3.5.6) escreve-se:

$$K(\mathbf{y}|\mathbf{x}, \lambda) = \prod_{k=1}^{d} \lambda^{1-w_k}(1 - \lambda)^{w_k}$$

Aitchinson e Aitken (1976), mostraram que para variáveis binárias o estimador de máxima verossimilhança de λ também não é satisfatório ($\hat{\lambda}_{MV} = 1$), logo, propuseram usar o método de verossimilhança modificado, baseado em $\prod_j f^{(r)}(x_r)$ (ver (3.5.4)), para determinar outro estimador mais adequado da janela. Hall (1981), demonstrou teoricamente que o estimador Jacknife pode ter um comportamento errático ainda para grandes amostras e além disso que é fortemente influenciado por categorias multinomiais vazias ou quase vazias. Por último Hall (1981) propôs um estimador alternativo para λ.

c) Dados Mistos: Contínuos e Binários

Dado o vetor aleatório misto $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$ Aitchinson e Aitken (1976), sugeriram usar um produto de funções Kernel tal como

$$K(\mathbf{w}|\mathbf{z}, \lambda_1, \lambda_2) = K_1(\mathbf{y}|\mathbf{z}, \lambda_1)\, K_2(\mathbf{x}|\mathbf{z}, \lambda_2)$$

onde K_1 é um Kernel multinormal como em (3.5.2) e K_2 é um Kernel binário como em (3.5.6). Logo, pode-se estimar a densidade do vetor misto, $f(\mathbf{w})$, por:

$$\hat{f}(\mathbf{w}) = \frac{1}{n} \sum_{j=1}^{n} K_1(\mathbf{y}|\mathbf{y}_j, \hat{\lambda}_1)\, K_2(\mathbf{x}|\mathbf{x}_j, \hat{\lambda}_2)$$

onde $\hat{\lambda}_1$ e $\hat{\lambda}_2$ são as janelas estimadas. O método de verossimilhança Jacknife para escolher λ_1 e λ_2 também é simples neste caso, a diferença com os anteriores é que aqui é necessário maximizar uma função de duas variáveis.

É importante observar que usar um produto de Kernels não implica de maneira alguma, em independência entre os componetes binários e contínuos de **w** (Seber, 1984 p. 323).

Aitchinson e Aitken (1976), também discutiram a extensão deste enfoque para variáveis categóricas com mais de duas categorias e propuseram duas soluções para este caso; a primeira é transformar as variáveis categóricas em binárias e a segunda, usar uma modificação do Kernel binário que permita levar em conta as categorias de todas as variáveis.

Os métodos Kernel também podem ser adaptados para trabalhar com valores perdidos (*missing value*) e/ou observações não classificadas, isto é, o chamado problema de misturas (*mixture-problem*) (ver Murray e Titterington (1978)). Outras extensões interessantes são os Kernel de janela variável ou Kernel adaptativos (*variable width or adaptative Kernels*), propostos pela primeira vez por Breimal *et al.* (1977), que são mais adequados para distribuições assimétricas como a lognormal, (ver também Habbema *et al.* 1978 e Remme *et al.* 1980 p.103).

3.5.2 - Discriminação Segundo o Vizinho Mais Próximo

Os métodos do vizinho mais próximo são bem conhecidos na Análise de Conglomerados porém, uma aplicação menos difundida deles é em problemas de discriminação. O artigo de Devroy e Wagner (1982), fornece uma descrição detalhada destes métodos para um caso geral, sem

considerar a estrutura particular das observações em presença de mistura de variáveis. No mesmo artigo os autores dão resultados muito interessantes sobre a convergência dos métodos.

Esta sub-seção estará baseada principalmente no artigo de Wojciechowski (1987), que construiu uma regra de classificação para misturas baseado nos métodos do vizinho mais próximo propostos por Cover e Hart (1967).

Seja π uma população geral de indivíduos caracterizados pelo vetor misto, (p + q) variado, $\mathbf{w}' = (\mathbf{x}', \mathbf{y}')$ que pertence ao espaço métrico (W, d), onde de maneira geral

$$\mathbf{W} = \{a_{11}, a_{12}, \ldots\} \times \ldots \times \{a_{q1}, a_{q2}, \ldots\} \times \mathbb{R}^p$$

e d é uma distância definida para quaisquer $\mathbf{w}_1$ e $\mathbf{w}_2$ por

$$d^2(\mathbf{w}_1, \mathbf{w}_2) = \sum_{j=1}^{p+q} d_j^2(w_{1j}, w_{2j})$$

onde $d_j^2(w_{1j}, w_{2j}) = (w_{1j} - w_{2j})^2$ se a j-ésima componente de $\mathbf{w}$ é contínua ou categórica ordinal e

$$d^2(w_{1j}, w_{2j}) = \begin{cases} 0 & \text{se} \quad w_{1j} = w_{2j} \\ 1 & \text{se} \quad w_{1j} \neq w_{2j} \end{cases}$$

se a j-ésima componente de $\mathbf{w}$ é categórica nominal (não ordenada).

Assumindo que a população π está dividida em dois grupos π_1 e π_2 e que tem-se uma amostra de treinamento de n indivíduos escolhidos ao acaso de π, tais que n_1 pertencem ao grupo π_1 e $n_2 = n - n_1$ pertencem ao grupo π_2 então, dada uma nova observação $\mathbf{w}$ que se sabe pertence a π mas não a qual dos dois grupos, necessita-se construir uma regra de

classificação que permita alocar w em π_1 ou π_2. Como não se dispõe de informação sobre a forma da distribuição de probabilidades de π, será usado um método não paramétrico chamado discriminação segundo o vizinho mais próximo que está baseado na definição dada a seguir.

Definição 3.5.1: Sejam w, w_1, ..., w_n valores do espaço métrico (W, d). O elemento $\tilde{w}_n \in \{w_1, ..., w_n\}$ é dito vizinho mais próximo de w entre w_1, ..., w_n se

$$d(w, \tilde{w}_n) = \min_{1 \leq j \leq n} d(w, w_j)$$

O elemento $\tilde{w}_n$ da definição anterior pode não estar determinado de maneira única e por esta razão é introduzido o seguinte suplemento à definição.

Seja z, z_1, ..., z_n uma sequência de valores gerados de uma distribuição uniforme em $[0,1]$. O elemento w_i é dito estar mais próximo de w que w_j se:

i) $d(w, w_i) < d(w, w_j)$ ou

ii) $d(w, w_i) = d(w, w_j)$ e $|z - z_i| < |z - z_j|$ ou

iii) $d(w, w_i) = d(w, w_j)$ e $|z - z_i| < |z - z_j|$ e $i < j$

logo, a regra de classificação segundo o vizinho mais próximo é :

Alocar w a π_r se o vizinho mais próximo de w,

$\tilde{w}_n$ pertence a π_r $\quad (r = 1, 2)$.

Uma extensão trivial desta regra é a classificação segundo os k vizinhos mais próximos que consiste em alocar a observação w no grupo ao qual pertencem a maioria destes vizinhos. Se o número de vizinhos é

igual nos dois grupos π_1 e π_2, então a decisão é alocar **w** no grupo ao qual pertence o vizinho mais próximo.

Cover e Hart (1967), demonstraram que o vizinho mais próximo $\tilde{\mathbf{w}}_n$ de **w** tende a **w** quase certamente quando o número de observações, n, cresce infinitamente ($n \to \infty$).

CAPÍTULO 4

ANÁLISE DE UM PROBLEMA PRÁTICO

4.1 - INTRODUÇÃO

Este capítulo trata de ilustrar os métodos que são objeto de principal interesse deste trabalho, conforme seção 3.1.

O critério de escolha do problema foi corresponder a um conjunto de dados com potencialidade de exibir desempenhos comparativos entre os métodos, e para um mesmo método, em mais de uma situação.

O referencial adotado para a estrutura dos dados foi o seguinte: buscar um problema cuja solução demandasse técnicas de discriminação, especificamente de Análise Discriminante, envolvendo grupos caracterizados por misturas de variáveis contínuas e categóricas sendo interessante uma situação de três grupos sobre os quais as mesmas variáveis sejam observadas e tais que se diferenciam ou não ora pelas contínuas, ora pelas categóricas com desejável interação entre ambos

os conjuntos. Neste contexto, dos problemas encontrados foi escolhido o exposto a seguir.

Trata-se de diagnosticar diferenças do perfil do aluno ingressante a três cursos aqui denominados A, B e C. Este perfil pretende tocar duas naturezas: acadêmica e sócio-econômica.

Sobre cada aluno ingressante está disponível informação relacionada ao perfil sócio-econômico obtida por questionário e também notas relativas ao desempenho em provas de seleção.

Para efeitos deste diagnóstico foram selecionadas como variáveis informativas do perfil sócio-econômico do aluno as seguintes variáveis:

X_1: Tipo de estabelecimento de ensino em que cursou o primeiro grau:
- Predominantemente público (0)
- Predominantemente particular (1)

X_2: Tipo de estabelecimento de ensino em que cursou o segundo grau:
- Predominantemente público (0)
- Predominantemente particular (1)

X_3: Curso de segundo grau concluído:
- Técnico (0)
- Comum (1)

X_4: Período em que cursou o segundo grau:
- Predominantemente diurno (0)
- Predominantemente noturno (1)

X_5: Categoria à qual pertence a ocupação do pai:
- Estrato superior (1)
- Estrato médio (2)
- Estrato inferior (3)

X_6: Participação na vida econômica da família:

- Não trabalha (0)
- Trabalha (1)

O perfil acadêmico fica representado pelas seguintes variáveis:

Port: Nota na prova de português,
Biol: Nota na prova de biologia,
Quim: Nota na prova de química,
Hist: Nota na prova de história,
Fis: Nota na prova de física,
Geog: Nota na prova de geografia,
Mat: Nota na prova de matemática e
L_est: Nota na prova de língua estrangeira

que assumem valores de zero a dez.

As variáveis disponíveis são de natureza mista desde que as notas podem ser consideradas contínuas e as demais são, quatro binárias e uma ternária (X_5, categórica ordinal com três categorias).

Por outro lado dispondo de três cursos, grupos, será possível mostrar distintas situações dentro do contexto deste trabalho ao comparar os grupos dois a dois.

A escolha final recaiu sobre este problema pois análises marginais sobre o grupo de contínuas e sobre o grupo das categóricas revelam distintos padrões de discriminação, ora só pelas variáveis contínuas, ora só pelas variáveis categóricas, ora por ambas. Tanto as análises preliminares como os dados estão documentados no Apêndice A.

A aplicação dos métodos aos dados é apresentada nas seções seguintes, ilustrando cada modelo de interesse, quando possível, na ordem de apresentação do capítulo 3.

Não foi necessário desenvolver programas computacionais para a aplicação destes métodos. A FLD de Fisher e as três modificações propostas por Vlachonikolis e Marriott (1982) (seção 2.3) podem ser calculadas usando o procedimento PROC DISCRIM do sistema de programa SAS. Para isto basta considerar como variáveis explanatórias, as variáveis contínuas e as categóricas para a FLD de Fisher, ou as variáveis contínuas e as novas variáveis (construídas a partir das originais) propostas para cada uma das modificações. Por outro lado, o programa para a aplicação do método baseado no modelo de posição (seção 3.2) foi cedido gentilmente pelo Dr. W. J. Krzanowski da Universidade de Exeter da Inglaterra.

Notando que no caso em estudo trata-se de oito variáveis contínuas, cinco binárias e uma ternária, na seção 4.2 são obtidos resultados só para o modelo de posição e a FLD de Fisher. As modificações da FLD estudadas na subseção 3.5.2 não são consideradas devido ao grande número de parâmetros necessários para construí-las.

Na seção 4.3 ilustra-se como a análise do problema beneficia-se de outras técnicas, explorando os métodos de discriminação para misturas após a redução de dimensão do conjunto das variáveis contínuas e do conjunto de variáveis categóricas.

4.2 - DISCRIMINAÇÃO COM O CONJUNTO COMPLETO DE VARIAVEIS

Nesta seção apresentam-se os resultados da aplicação do método de dicriminação baseado no modelo de posição e da FLD de Fisher aos dados documentados no Apêndice A, tomando os cursos dois a dois.

As modificações da FLD (seção 3.2.5) não serão calculadas pois o número de parâmetros necessários para construí-las é muito grande comparado com os tamanhos de amostra disponíveis. Fazendo as

adaptações necessárias para construir as FLD modificadas com as 8 variáveis contínuas, 4 binárias e a ternária X_5: ocupação do pai[1], o número de parâmetros que deveria ser estimado para construí-las é:

104 para a Primeira Modificação da FLD,
864 para a Segunda Modificação da FLD,
252 para a Terceira Modificação da FLD,

e os tamanhos de amostra disponíveis são: 70 no curso A, 90 no curso B e 31 no curso C.

4.2.1 - Análise Discriminate para os Cursos A e B

a) Modelo de Posição

- Variáveis consideradas:

  8 contínuas, 5 binárias e 1 ternária.

  A variável ternária é substituída por duas binárias (b_1, b_2) conforme visto na sub-seção 3.2.3.

- Número de categorias multinomiais geradas pelas variáveis categóricas no modelo:

  $2^5*(3) = 96$ categorias

- Número de observações nos grupos (cursos):

  Curso A: 70
  Curso B: 90

- Categorias com maior probabilidade de ocorrência estimada a partir

---

(1) Sao necessárias algumas adaptaçoes pois as modificaçoes foram propostas para misturas de variáveis contínuas e binárias. Na seçao 4.4 sao descritas estas adaptaçoes.

de um modelo log-linear de primeira ordem (não foi possível usar um modelo de segunda ordem pois as tabelas de contingência observadas são muito esparsas).

Tabela 4.2.1 - Descrição das categorias com maior probabilidade de ocorrência no curso A.

71: 0110001 (1)	69: 0010001	87: 0110101
4,78% (2)	4,26%	3,19%
1o grau público (3)	1o grau público	1o grau público
2o grau particular	2o grau público	2o grau partic.
2o grau comum	2o grau comum	2o grau comum
2o grau diurno	2o grau diurno	2o grau diurno
Oc. pai: e. médio	Oc. pai: e. médio	Oc. pai: médio
Não trabalha	Não trabalha	Trabalha

(1) Número de categoria e valores de variáveis binárias.

(2) Probabilidade de ocorrência da categoria em percentagem.

(3) Descriçao da categoria.

Na tabela anterior observa-se as seguintes características comuns às três categorias com maior probabilidade de ocorrência no curso A:

- estabelecimento de ensino predominantemente público no primeiro grau.
- curso comum (não técnico) em período noturno no segundo grau.
- ocupação do pai no estrato médio.

Estas categorias representam apenas 12,23% da probabilidade de observar alguma das 96 categorias possíveis portanto, não é recomendável construir um perfil representativo dos alunos do curso A baseados somente nesta informação.

Tabela 4.2.2 - Descrição das categorias com maior probabilidade de ocorrência no curso B.

40: 1110010[1]	72: 1110001	39: 0110010
21,57%[2]	14,54%	7,40%
1º grau partic.[3]	1º grau partic.	1º grau público
2º grau partic.	2º grau partic.	2º grau partic.
2º grau comum	2º grau comum	2º grau comum
2º grau diurno	2º grau diurno	2º grau diurno
Oc. pai: sup.	Oc. pai: médio	Oc. pai: sup.
Não trabalha	Não trabalha	Não trabalha

(1) Número de categoria e valores de variáveis binárias.

(2) Probabilidade de ocorrência da categoria (percentagem).

(3) Descriçao da categoria.

A tabela anterior mostra que no curso B, as três categorias com maior probabilidade de ocorrência representam 43,51% da probabilidade de observar alguma das 96 categorias geradas pelas variáveis informativas do aspecto sócio-econômico, isto sugere uma caracterização dos alunos do curso B considerando os elementos comuns nessa três categorias. Estes são:

- Segundo grau em estabelecimento prodominantemente particulares, curso comum e em período diurno.
- Ocupação do pai: estrato superior e médio.
- Não participa da vida econômica da família.

- Análise das médias das variáveis contínuas estimadas por regressão multivariada.

As estimativas das médias das notas são obtidas a partir do seguinte modelo de regressão multivariada:

$$\begin{bmatrix} Port_i \\ Biol_i \\ \vdots \\ Mat_i \\ L\_est_i \end{bmatrix} = \begin{bmatrix} \nu_i^{(Port)} & \alpha_{i1}^{(Port)} & \cdots & \alpha_{i7}^{(Port)} \\ \nu_i^{(Biol)} & \alpha_{i1}^{(Biol)} & \cdots & \alpha_{i7}^{(Biol)} \\ \vdots & \vdots & & \vdots \\ \nu_i^{(Mat)} & \alpha_{i1}^{(Mat)} & \cdots & \alpha_{i7}^{(Mat)} \\ \nu_i^{(L\_est)} & \alpha_{i1}^{(L\_est)} & \cdots & \alpha_{i7}^{(L\_est)} \end{bmatrix} \begin{bmatrix} 1 \\ X_1 \\ \vdots \\ b_1 \\ b_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_i^{(Port)} \\ \varepsilon_i^{(Biol)} \\ \vdots \\ \varepsilon_i^{(Mat)} \\ \varepsilon_i^{(L\_est)} \end{bmatrix}$$

$$i = 1, 2 \qquad (4.2.1)$$

onde b_1 e b_2 são as variáveis binárias pelas quais foi substituída a ternária X_5.

Observe-se que no modelo (4.2.1) somente foram considerados os efeitos principais das variáveis categóricas por razões de consistência com o modelo log-linear usado para estimar as probabilidades multinomiais.

Depois de estimar os parâmetros de (4.2.1) tem-se os seguintes modelos para as médias das variáveis contínuas do curso A:

$$Port_A = 3{,}31 + 0{,}12\ X_1 + 0{,}48\ X_2 + 0{,}35\ X_3 - 0{,}74\ X_4 + 0{,}29\ X_6 - 0{,}02\ b_1 - 0{,}11\ b_2$$

$$Biol_A = 2{,}07 + 0{,}40\ X_1 + 0{,}19\ X_2 + 0{,}40\ X_3 - 0{,}46\ X_4 + 0{,}39\ X_6 - 0{,}32\ b_1 - 0{,}38\ b_2$$

$$Quim_A = 2,55 + 0,72\ X_1 + 0,15\ X_2 + 0,66\ X_3 - 0,78\ X_4 + 1,12\ X_6 + \\ + 0,10\ b_1 + 0,14\ b_2$$

$$Hist_A = 4,06 - 0,01\ X_1 + 0,30\ X_2 + 0,02\ X_3 - 0,78\ X_4 + 0,47\ X_6 - \\ - 0,18\ b_1 - 0,27\ b_2$$

$$Fis_A = 2,31 + 0,32\ X_1 + 0,68\ X_2 + 0,11\ X_3 + 0,08\ X_4 + 0,27\ X_6 - \\ - 1,01\ b_1 - 0,37\ b_2$$

$$Geo_A = 4,14 + 0,26\ X_1 - 0,04\ X_2 + 0,21\ X_3 - 0,36\ X_4 + 0,36\ X_6 - \\ - 0,36\ b_1 - 0,20\ b_2$$

$$Mat_A = 3,42 + 0,28\ X_1 + 0,42\ X_2 + 0,09\ X_3 - 0,11\ X_4 + 0,55\ X_6 - \\ - 0,03\ b_1 + 0,11\ b_2$$

$$L\_est_A = 2,15 + 1,27\ X_1 + 0,26\ X_2 + 0,70\ X_3 + 0,38\ X_4 + 0,76\ X_6 + \\ + 1,47\ b_1 + 0,78\ b_2$$

As equações anteriores mostram que, em geral a contribuição das variáveis categóricas às médias condicionais das notas é pequena comparada com o termo independente. Em todos os casos estas contribuições são menores que um desvio padrão da média correspondente (veja tabela A.1). É interessante notar que na variável L_est (língua estrangeira) o efeito conjunto de $X_1 = 1$ e $b_1 = 1$, que representam 1º grau em escola particular e ocupação do pai no estrato superior, modifica consideravelmente o termo independente.

Os modelos estimados para as médias das nota dos curso B são:

$$Port_B = 5,52 + 0,35\ X_1 + 0,06\ X_2 + 0,12\ X_3 - 0,09\ X_4 + 0,21\ X_6 + \\ + 0,06\ b_1 - 0,12\ b_2$$

$$Biol_B = 5,56 - 0,01\ X_1 + 0,18\ X_2 + 0,23\ X_3 - 0,67\ X_4 - 0,46\ X_6 - \\ - 0,03\ b_1 - 0,31\ b_2$$

$$Quim_B = 7,62 - 0,02\ X_1 + 0,69\ X_2 - 0,92\ X_3 + 1,05\ X_4 - 0,93\ X_6 + \\ + 0,41\ b_1 - 0,07\ b_2$$

$$Hist_B = 6,78 + 0,21\ X_1 - 0,15\ X_2 - 0,06\ X_3 - 0,78\ X_4 - 0,16\ X_6 + \\ + 0,22\ b_1 - 0,01\ b_2$$

$$Fis_B = 7,22 - 0,16\ X_1 + 0,51\ X_2 + 0,03\ X_3 + 1,77\ X_4 + 0,63\ X_6 - \\ - 0,13\ b_1 - 0,05\ b_2$$

$$Geo_B = 6,51 - 0,01\ X_1 - 0,26\ X_2 + 0,08\ X_3 - 0,48\ X_4 - 0,24\ X_6 - \\ - 0,11\ b_1 - 0,24\ b_2$$

$$Mat_B = 7,97 - 0,23\ X_1 + 0,52\ X_2 - 0,62\ X_3 + 0,69\ X_4 - 0,76\ X_6 - \\ - 0,35\ b_1 - 0,52\ b_2$$

$$L\_est_B = 7,70 + 0,12\ X_1 + 0,29\ X_2 + 0,20\ X_3 - 0,72\ X_4 + 0,95\ X_6 + \\ + 0,36\ b_1 - 0,23\ b_2$$

Neste caso, como no curso A, a influência dos efeitos principais das variáveis categóricas é pequena comparada com o termo independente dos modelos estimados para as médias das notas.

A tabela 4.2.3, que se apresenta a seguir, mostra as estimativas das médias das variáveis contínuas nas categorias com maior probabilidade de ocorrência nos cursos A e B. Estas estimativas são obtidas substituindo nos modelos acima, os valores que assumem as variáveis binárias nas categorias de interesse. Na tabela, observa-se que o desempenho geral dos alunos do curso B é consideravelmente melhor que o desempenho dos alunos do curso A, em todos os casos as médias do curso B são maiores que as médias do curso A. Nas seis categorias consideradas, as maiores diferenças no desempenho foram observadas em física.

Tabela 4.2.3 - Estimativa das médias das notas nas categorias com maior probabilidade de ocorrência nos cursos A e B.

	Matérias							
Categ	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_es
71	4,04[1]	2,29	3,51	4,11	2,73	4,12	4,04	3,9
	5,59[2]	5,79	7,33	6,56	7,71	6,09	7,35	7,9
69	3,56	2,10	3,35	3,81	2,05	4,16	3,62	3,64
	5,52	5,61	6,64	6,71	7,20	6,35	6,82	7,66
87	4,33	2,68	4,62	4,58	3,00	4,47	4,58	4,66
	5,80	5,32	6,40	6,40	7,07	5,84	6,58	8,90
40	4,25	2,74	4,19	4,18	2,41	4,22	4,17	5,86
	6,11	6,06	7,78	7,00	7,46	6,20	7,29	8,67
72	4,16	2,69	4,23	4,10	3,05	4,38	4,32	5,17
	5,94	5,78	7,31	6,78	7,55	6,07	7,12	8,07
39	4,13	2,34	3,47	4,20	2,09	3,96	3,89	4,59
	5,76	6,07	7,80	6,79	7,62	6,22	7,51	8,55

(1) Média no curso A

(2) Média no curso B

- Matriz de resíduos que estima a matriz de variância-covariância das variáveis contínuas dadas as categorias, definida em (3.2.16).

	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_est
Port	0,93							
Biol	0,14	0,98						
Quim	0,26	0,41	1,30					
Hist	0,29	0,15	0,12	0,69				
Fis	0,19	0,42	0,79	0,23	1,95			
Geog	0,21	0,15	0,04	0,36	0,06	0,62		
Mat	0,04	0,02	0,34	0,04	0,87	-0,12	1,21	
L_est	0,38	0,01	0,47	0,40	0,53	0,35	0,28	2,20

(4.2.2)

- Funções de discriminação conforme a expressão (3.2.8) para as categorias com maior probabilidade de ocorrência nos grupos A e B, construídas a partir das estimativas das probabilidades multinomiais nas tabelas 4.2.1 e 4.2.2, as estimativas das médias das notas na tabela 4.3.2 e a matriz de dispersão estimada (4.2.2).

Categoria 71:

49,08 + 0,30 Port - 2,58 Biol - 1,03 Quim - 1,30 Hist - 0,19 Fis -
- 1,80 Geog - 2,23 Mat - 0,80 L_est

Categoria 69:

49,37 - 0,16 Port - 2,59 Biol - 0,33 Quim - 1,89 Hist - 0,56 Fis -
- 1,69 Geog - 2,05 Mat - 0,71 L_est

Categoria 87:

33,19 - 0,33 Port - 2,38 Biol + 0,92 Quim - 0,62 Hist - 1,09 Fis -
- 0,41 Geog - 0,75 Mat - 1,52 L_est

Categoria 40:

48,05 - 0,15 Port - 2,09 Biol - 0,99 Quim - 2,25 Hist - 0,55 Fis -
- 1,61 Geog - 1,95 Mat + 0,02 L_est

Categoria 72:

43,39 - 0,22 Port - 2,15 Biol - 0,60 Quim - 2,33 Hist - 0,43 Fis -
- 0,90 Geog - 1,75 Mat - 0,24 L_est

Categoria 39:

54,10 - 0,38 Port - 2,51 Biol - 1,42 Quim - 1,22 Hist - 0,31 Fis -
- 2,50 Geog - 2,43 Mat - 0,54 L_est

Chamando $\xi(m)$ à função de discriminação na categoria m tem-se que a regra de classificação nesta categoria é:

Alocar no curso A se $\xi(m) \geq 0$
caso contrário alocar no curso B.

(4.2.3)

- Taxas de erro estimadas usando o método de Lachenbruch[2]

As proporções estimadas de observações mal classificadas pela regra (4.2.3) são:

Curso A: 4,29%
Curso B: 1,11%
Erro total: 2,7%.[2]

---

(1) O método da Lachenbruch (ou 'leaving-one-out'), foi escolhido para estimar as taxas de erro porque reduz o vício do método de resubstituçao.

(2) Erro total calculado, considerando probabilidades prévias dos cursos igual a 0,5.

b) Função Linear Discriminante de Fisher

- Número de variáveis consideradas: 14

  (Note-se que para calcular a FLD não é necessário especificar o tipo das variáveis incluídas).

- A regra de classificação baseada na FLD de Fisher para os cursos A e B é a seguinte:

  Alocar no curso A se:

$$39{,}02 - 0{,}42\,\text{Port} - 2{,}00\,\text{Biol} - 0{,}25\,\text{Quim} - 1{,}73\,\text{Hist} - 0{,}68\,\text{Fis} - 0{,}94\,\text{Geog} - 1{,}50\,\text{Mat} - 0{,}55\,\text{L\_est} - 0{,}22\,X_1 + 1{,}31\,X_2 + 1{,}02\,X_3 - 1{,}66\,X_4 + 0{,}20\,X_5 + 2{,}83\,X_6 \geq 0$$

  e no curso B caso contrário. (4.2.4)

- Taxas de erro estimadas usando o método de Lachenbruch.

As proporções de má classificação estimadas para a regra (4.2.4) são:

Curso A: 4,29%

Curso B: 0,00%

Erro total: 2,14%.

4.2.2 - Análise Discriminante para cursos A e C

a) Modelo de Posição

- Variáveis consideradas:

  8 contínuas, 5 binárias e 1 ternária

  (substituída por duas binárias conforme sub-seção 3.2.3).

- Número de categorias multinomiais: 96

- Número de observações nos grupos:

  Curso A: 70

  Curso B: 31

- Categorias com maior probabilidade de ocorrência estimadas a partir de um modelo log-linear de primeira ordem.

A tabela 4.2.1 mostra as categorias mais prováveis de serem observadas no curso A portanto, aqui será analisado somente o curso C.

Tabela 4.2.4 - Descrição das categorias com maior probabilidade de ocorrência no curso C.

87: 0110101[1]	71: 0110001	55: 0110110
4,57%[2]	4,28%	4,22%
1º grau público[3]	1º grau público	1º grau público
2º grau partic.	2º grau partic.	2º grau partic.
2º grau comum	2º grau comum	2º grau comum
2º grau diurno	2º grau diurno	2º grau diurno
Oc. pai: médio	Oc. pai: médio	Oc. pai: super.
Trabalha	Não Trabalha	Trabalha

(1) Número de categoria e valores de variáveis binárias.

(2) Probabilidade de ocorrência da categoria (percentagem).

(3) Descriçao da categoria.

As categorias anteriores representam apenas 13,07% da probabilidade de observar alguma das 96 categorias possíveis portanto neste caso, como para o curso A, não se consegue detectar um perfil dominante dos alunos do curso C levando em conta somente estas categorias.

É interessante comentar que nas categorias 87, 71 e 55 as variáveis que representam escolaridade assumem os mesmos valores:

primeiro grau público, segundo grau particular em curso comum e em período diurno.

- Análise das médias das variáveis contínuas estimadas por regressão multivariada

O modelo de regressão utilizado neste caso é análogo ao modelo (4.2.1) pois as probabilidades multinomiais foram estimadas usando um modelo log-linear de primeira ordem, como no caso anterior.

Os modelos estimados para as médias das notas do curso A foram apresentadas na seção anterior. A seguir serão apresentados os modelos correspondentes para o curso C.

$$Port_C = 5{,}54 + 0{,}42\,X_1 - 0{,}46\,X_2 - 0{,}01\,X_3 + 0{,}52\,X_4 + 0{,}39\,X_6 - 0{,}21\,b_1 + 0{,}27\,b_2$$

$$Biol_C = 5{,}04 + 0{,}09\,X_1 - 0{,}01\,X_2 + 0{,}36\,X_3 + 1{,}46\,X_4 - 1{,}03\,X_6 + 0{,}37\,b_1 + 0{,}60\,b_2$$

$$Quim_C = 6{,}21 + 0{,}16\,X_1 + 0{,}24\,X_2 + 0{,}71\,X_3 + 0{,}86\,X_4 - 0{,}28\,X_6 + 0{,}24\,b_1 + 0{,}14\,b_2$$

$$Hist_C = 6{,}12 + 0{,}27\,X_1 + 0{,}30\,X_2 + 0{,}55\,X_3 + 0{,}63\,X_4 + 0{,}13\,X_6 - 0{,}43\,b_1 - 0{,}11\,b_2$$

$$Fis_C = 6{,}28 - 0{,}07\,X_1 + 0{,}75\,X_2 - 0{,}46\,X_3 + 0{,}54\,X_4 - 0{,}91\,X_6 - 0{,}14\,b_1 + 0{,}37\,b_2$$

$$Geo_C = 6{,}20 + 0{,}49\,X_1 - 0{,}60\,X_2 - 0{,}38\,X_3 + 0{,}12\,X_4 - 0{,}42\,X_6 + 0{,}45\,b_1 + 0{,}97\,b_2$$

$$Mat_C = 7,85 + 0,08\ X_1 - 0,76\ X_2 - 0,43\ X_3 - 1,25\ X_4 - 0,10\ X_6 - 1,43\ b_1 - 0,87\ b_2$$

$$L\_est = 7,91 + 0,93\ X_1 + 0,57\ X_2 - 0,48\ X_3 - 0,21\ X_4 + 0,59\ X_6 - 0,44\ b_1 - 0,89\ b_2$$

As equações anteriores mostram que em relação ao termo independente, a contribuição das variáveis binárias às médias condicionais das notas é pequena sendo quase sempre menor que um desvio padrão da média correspondente (ver tabela A.1).

Tabela 4.2.5 - Estimativas das médias das notas nas categorias com maior probabilidade de ocorrência nos cursos A e C.

	Matérias							
Categ	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_est
71	4,03[1]	2,29	3,51	4,11	2,73	4,12	4,04	3,90
	5,34[2]	5,98	7,31	6,85	6,94	6,21	7,79	7,12
87	4,33	2,68	4,62	4,58	3,00	4,47	4,58	4,66
	5,73	4,95	7,03	6,98	6,03	5,79	5,68	7,71
69	3,56	2,10	3,35	3,81	2,05	4,16	3,62	3,64
	5,80	5,99	7,06	6,55	6,19	6,79	6,55	6,54
55	4,42	2,73	4,58	4,67	2,35	4,32	4,44	5,35
	5,25	4,72	7,12	6,67	5,51	5,27	5,12	8,16

(1) Média no curso A

(2) Média no curso C

Na tabela 4.2.5 observa-se que em todos os casos as médias das notas no

curso A são menores que no curso C. As maiores diferenças foram observadas em física.

- Matriz de variância e covariância condicional estimada conforme (3.2.16).

	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_est
Port	1,06							
Biol	0,13	0,89						
Quim	0,29	0,46	1,66					
Hist	0,38	0,24	0,13	0,84				
Fis	0,34	0,50	1,05	0,40	1,96			
Geog	0,31	0,11	0,02	0,54	0,23	0,78		
Mat	0,11	0,20	0,45	0,06	0,73	-0,04	0,92	
L_est	0,49	0,08	0,61	0,62	0,53	0,53	0,12	2,63

(4.2.5)

- Funções de discriminação nas categorias mais prováveis de ser observadas nos cursos A e C, calculadas usando as estimativas das probabilidades multinomiais nas tabelas 4.2.1 e 4.2.4, as estimativas das médias das notas na tabela 4.2.5 e a matriz de dispersão estimada (4.2.5).

Categoria 71:

34,64 + 0,54 Port - 2,89 Biol - 1,05 Quim - 1,24 Hist - 0,23 Fis -
- 1,32 Geog - 0,56 Mat - 0.36 L_est

Categoria 87:

23,26 - 0,05 Port - 1,55 Biol - 0,40 Quim - 1,98 Hist - 0,37 Fis +
+ 0,40 Geog - 0,14 Mat - 0.54 L_est

Categoria 69:

$$47{,}38 - 0{,}47\ \text{Port} - 3{,}18\ \text{Biol} - 0{,}88\ \text{Quim} - 0{,}24\ \text{Hist} + 0{,}61\ \text{Fis} - 2{,}79\ \text{Geog} - 2{,}57\ \text{Mat} - 0.10\ \text{L\_est}$$

Categoria 55:

$$15{,}73 + 0{,}43\ \text{Port} - 1{,}21\ \text{Biol} - 0{,}56\ \text{Quim} - 1{,}80\ \text{Hist} - 0{,}89\ \text{Fis} + 0{,}71\ \text{Geog} + 0{,}67\ \text{Mat} - 0.54\ \text{L\_est}$$

Chamando $\xi(m)$ à função de discriminação na categoria m, a regra de classificação para observações que pertencem a esta categoria é:

Alocar no curso A se $\xi(m) \geq 0$
caso contrário alocar no curso C.

(4.2.6)

- Taxas de erro estimadas usando o método de Lachenbruch, para a regra de classificação (4.2.6).

  Curso A: 5,71%
  Curso C: 6,45%
  Erro total: 6,08%

b) Funcão Linear Discriminante de Fisher

- Número de variáveis consideradas: 14
- A regra de classificação baseada na FLD de Fisher para os cursos A e C é a seguinte:

  Alocar no curso A se:

$$24{,}09 - 0{,}19\ \text{Port} - 1{,}79\ \text{Biol} - 0{,}35\ \text{Quim} - 0.76\ \text{Hist} - 0{,}30\ \text{Fis} - 0{,}95\ \text{Mat} - 0{,}56\ \text{L\_est} + 2{,}67\ X_1 + 0{,}20\ X_2 + 1{,}81\ X_3 - 0{,}15\ X_4 + 0{,}93\ X_5 + 0{,}93\ X_6 \geq 0$$

(4.3.7)

alocar no curso C caso contrário.

- Taxas de erro estimadas usando o método de Lachenbruch, para a regra de classificação (4.2.7).

  Curso A: 8,57%

  Curso C: 3,23%

  Erro total: 5,9%.

4.2.3 - Análise Discriminante para os cursos B e C

a) Modelo de Posição

- Variáveis consideradas:

  8 contínuas, 5 binárias e uma ternária.

  (substituída por duas binárias conforme sub-seção 3.2.3).

- Número de categorias multinomiais: 96

- Número de observações por cursos:

  Curso B: 90

  Curso C: 31

- As categorias com maior probabilidade de ocorrência para os cursos B e C foram apresentadas nas tabelas 4.2.2 e 4.2.4, respectivamente portanto não serão repetidas aqui.

- Comparação das médias das notas nas categorias mais prováveis para os cursos B e C.

Na tabela seguinte observa-se que o desempenho médio dos alunos dos cursos B e C foi similar em todas as matérias consideradas na avaliação. As médias em português, física e língua estrangeira, nas 6 categorias consideradas, foram ligeiramente maiores no curso B porém, estas diferenças não são consideráveis.

Tabela 4.2.6 - Médias das notas nas categorias mais prováveis nos cursos B e C.

	Matérias							
Categ	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_est
40	6,11[1]	6,06	7,78	7,00	7,46	6,20	7,29	8,67
	5,28[2]	5,85	7,56	6,83	6,49	6,18	5,30	8,50
72	5,94	5,78	7,31	6,78	7,55	6,07	7,12	8,07
	5,76	6,08	7,47	7,11	7,01	6,70	5,86	8,05
39	5,76	6,07	7,80	6,79	7,62	6,22	7,51	8,55
	4,86	5,76	7,40	6,54	6,42	5,69	5,23	7,57
87	5,80	5,32	6,40	6,40	7,07	5,84	6,58	8,90
	5,73	4,95	7,03	6,98	6,03	5,79	5,68	7,71
71	5,59	5,79	7,33	6,56	7,71	6,09	7,35	7,95
	5,34	5,98	7,31	6,85	6,94	6,21	7,79	7,12
55	5,97	5,60	6,88	6,63	6,99	5,98	6,75	9,50
	5,25	4,72	7,12	6,67	5,51	5,27	5,12	8,16

(1) Média no curso B

(2) Média no curso C

- Matriz de variância-covariância condicional estimada conforme (3.2.16):

	Port	Biol	Quim	Hist	Fis	Geo	Mat	L_est
Port	0,77							
Biol	0,13	1,20						
Quim	-0,01	0,38	0,69					
Hist	0,12	0,09	0,02	0,38				
Fis	-0,18	0,41	0,37	-0,01	1,68			
Geog	0,07	0,09	-0,01	0,11	-0,14	0,37		
Mat	-0,03	0,10	0,25	-0,01	0,89	-0,27	1,48	
L_est	0,04	-0,12	0,08	0,08	-0,03	0,12	0,00	1,36

(4.2.8)

- Funções de discriminação nas categorias mais prováveis de ser observadas nos cursos B e C, calculadas usando as estimativas das probabilidades mutinomiais nas tabelas 4.2.2 e 4.2.4, as estimativas das médias das notas na tabela 4.2.6 e a matriz de dispersão estimada (4.2.8).

Categoria 40:

$$-18{,}26 + 1{,}07\ \text{Port} - 0{,}09\ \text{Biol} - 0{,}17\ \text{Quim} - 0{,}04\ \text{Hist} + 0{,}00\ \text{Fis} + 1{,}04\ \text{Geog} + 1{,}56\ \text{Mat} + 0{,}00\ \text{L\_est}$$

Categoria 72:

$$8{,}75 + 0{,}45\ \text{Port} - 0{,}04\ \text{Biol} - 0{,}51\ \text{Quim} - 0{,}69\ \text{Hist} - 0.01\ \text{Fis} - 1{,}11\ \text{Geog} + 1{,}56\ \text{Mat} + 0{,}16\ \text{L\_est}$$

Categoria 39:

$$-34{,}22 + 1{,}09\ \text{Port} - 0{,}15\ \text{Biol} - 0{,}08\ \text{Quim} - 0{,}50\ \text{Hist} + 0.03\ \text{Fis} + 2{,}79\ \text{Geog} + 2{,}08\ \text{Mat} + 0{,}47\ \text{L\_est}$$

Categoria 87:

$$- 2{,}26 + 0{,}28\ \text{Port} + 0{,}87\ \text{Biol} - 2{,}02\ \text{Quim} - 2{,}19\ \text{Hist} + 0.61\ \text{Fis} + 0{,}85\ \text{Geog} + 0{,}67\ \text{Mat} + 1{,}13\ \text{L\_est}$$

Categoria 71:

$$- 8{,}15 + 0{,}47\ \text{Port} - 0{,}09\ \text{Biol} - 0{,}41\ \text{Quim} - 1{,}15\ \text{Hist} + 0.02\ \text{Fis} + 0{,}65\ \text{Geog} + 1{,}24\ \text{Mat} + 0{,}63\ \text{L\_est}$$

Categoria 55:

$$- 26{,}94 + 0{,}90\ \text{Port} + 0{,}81\ \text{Biol} - 1{,}69\ \text{Quim} - 1{,}54\ \text{Hist} + 0.61\ \text{Fis} + 2{,}99\ \text{Geog} + 1{,}51\ \text{Mat} + 0{,}97\ \text{L\_est}$$

Chamando $\xi(m)$ à função de discriminação na m-ésima categoria tem-se que, a regra para classificar observações desta categoria é:

Alocar no curso B se $\xi(m) \geq 0$
caso contrário alocar no curso C.

(4.2.9)

- Taxas de erro estimadas usando o método de Lachenbruch, para a regra de classificação (4.2.9).

  Curso B: 21,11%
  Curso C: 51,62%
  Erro total: 36,36%

b) Função Discriminante de Fisher

- Número de variáveis consideradas: 14
- Regra de classificação baseada na FLD de Fisher para os cursos B e C.

Alocar no curso B se:

$$-8{,}49 + 0{,}40\,\text{Port} - 0{,}04\,\text{Biol} - 0{,}18\,\text{Quim} + 0{,}10\,\text{Hist} + 0{,}44\,\text{Fis} - 0{,}02\,\text{Geog} + 0{,}48\,\text{Mat} + 0{,}23\,\text{L\_est} + 1{,}66\,X_1 - 0{,}60\,X_2 - 0{,}35\,X_3 - 1{,}49\,X_4 - 0{,}15\,X_5 - 2{,}9\,X_6 \geq 0$$

alocar no curso C caso contrário. (4.2.10)

- Taxas de erro estimadas usando o método de Lachenbruch.

As proporções estimadas de observações mal classificadas pela regra de alocação (4.2.10) são:

Curso B: 18,89%

Curso C: 38,71%

Erro Total: 28,8%.

4.2.4 - Alguns Comentários Finais sobre os Resultados

As análises apresentados nesta seção sugerem os seguintes comentários sobre as diferenças no perfil sócio-econômico e acadêmico dos alunos do curso A, B e C.

- Os alunos do curso A tiveram o pior desempenho geral. Esta característica diferencia claramente o curso A dos outros dois cursos e pode ser considerada a principal 'fonte' de discriminação para o curso A.
- Os alunos do curso B e C apresentaram desempenhos similares em todas as disciplinas avaliadas, isto dificulta a discriminação entre estes cursos. Esta dificuldade se reflete nas altas taxas de erro obtidas na discriminação de B e C.
- O desempenho geral dos alunos dos cursos B e C foi bom com notas médias quase sempre acima da nota mínima aprovatória, 5. No entanto, o desempenho geral no curso A pode ser considerado de regular a baixo.

- Os modelos estimados para as médias das variáveis contínuas mostram que, nos três cursos, a contribuição das variáveis categóricas é pequena comparada com o termo constante de cada modelo. Isto pode ser interpretado como uma indicação de que as médias correspondentes a cada disciplina têm pouca variabilidade entre as categorias do mesmo curso.

- As três categorias com maior probabilidade de ocorrência no curso B representam 43,51% da probabilidade de observar alguma das 96 categorias possíveis. Isto desperta interesse no perfil sócio-econômico descrito para este grupo baseado nas características comuns às três categorias. Este perfil é o seguinte:

  Grande parte dos alunos do curso B estudaram o segundo grau predominantemente em escolas particulares que oferecem cursos comuns e no período diurno. Em relação à ocupação do pai, predominam as ocupações consideradas de estrato médio e superior. Por último, uma grande proporção de alunos deste curso não trabalha, ou seja, não contribui à vida econômica da família.

- Não foi possível diferenciar adequadamente o perfil sócio-econômico dos alunos dos cursos A e C a partir dos resultados obtidos da aplicação da FLD e do método baseado no modelo de posição. No entanto, as análises univariadas das variáveis categóricas (tabela A.2 - A.7, do Apêndice A) mostram perfis muito parecidos nas variáveis de escolaridade (X_1, X_2, X_3 e X_4) e diferenças nas variáveis que representam o aspecto econômico (X_5 e X_6).

- As taxas de erro obtidas com o método baseado no modelo de posição e a FLD foram similares na discriminação dos cursos A e B, e A e C. Nestes dois casos a taxa de erro de classificação obtida com a FLD foi um pouco menor.

- No que diz respeito à discriminação de B e C as taxas de erro de classificação foram altas para os dois métodos considerados, sendo que a menor taxa de erro foi para a FLD.

4.3 - DISCRIMINAÇÃO COM REDUÇÃO DE DIMENSÃO E SELEÇÃO DE VARIAVEIS

Na seção anterior ficou clara a necessidade de reduzir a dimensão do problema antes de aplicar técnicas apropriadas para misturas. O excessivo número de categorias multinomiais (96), em relação a tamanhos de amostra moderadas (70, 90 e 31 para os cursos A, B e C, respectivamente), impediu a construção das funções lineares discriminantes modificadas propostas por Vlachonikolis e Marriott (1982) (ver seção 3.3) e dificultou a interpretação dos resultados obtidos usando o método baseado no modelo de posição (ver seção 3.2). No modelo de posição, muitas categorias não foram observadas ou o número de indivíduos observado foi muito pequeno, isto tornou impossível aproveitar a informação disponível para caracterizar as subpopulações (categorias) ou para compará-las mutuamente dentro e entre os cursos.

Nesta seção trata-se de ilustrar como uma redução adequada de variáveis contribui na interpretabilidade dos resultados e possibilita a aplicação dos distintos métodos de interesse. Para efeito de ilustração foi tratado o problema com os cursos A e C, com eles realizou-se novas análises que começaram pela redução do número de variáveis consideradas na discriminação.

O conjunto das notas, variáveis contínuas, foi reduzido usando o método de Componentes Principais. Foram mantidas as duas componentes derivadas a partir da matriz de correlação obtida através da matriz de variância e covariância total, da parte contínua dos cursos A e C, utilizando o procedimento PROC PRINCOMP do sistema de programas SAS.

Por outro lado, a escolha de um conjunto menor de variáveis categóricas foi feita usando o procedimento de seleção de variáveis para análise discriminante, através do PROC STEPDISC, do sistema de programas SAS.

4.3.1 - Redução da Dimensão do Problema

- Resultado da Análise de Componentes Principais

As duas primeiras componentes principais derivadas das 101 observações dos cursos A e C são as seguintes:

$$Prin1 = 0,34\ Port + 0,37\ Biol + 0,37\ Quim + 0,38\ Hist + 0,37\ Fis + 0,35\ Geog + 0,32\ Mat + 0,34\ L\_est$$

$$Prin2 = -0,29\ Port + 0,15\ Biol + 0,29\ Quim - 0,31\ Hist + 0,35\ Fis - 0,47\ Geog + 0,56\ Mat - 0,25\ L\_est$$

Prin1 representa um fator que se pode chamar de desempenho geral e explica 68,8% da variabilidade total dos dados.

Prin2 contrasta disciplinas de exatas com disciplinas de humanas e representa 9,4% da variabilidade das observações dos cursos A e C.

No total Prin1 e Prin2 explicam 78,2% da variabilidade.

- Resultado da seleção de variáveis categóricas

As variáveis selecionadas pelos métodos *stepwise*, *forward* e *backward* (opções do PROC STEPDISC do SAS), usando um nível de significância de 0,15 para que uma variável entre ou fique no modelo, coincidem e são as seguintes:

X_5: categoria à qual pertence a ocupação do pai
X_6: Participação na vida econômica da família.

Observe-se que esta escolha indica que as diferenças nos aspectos econômicos respondem mais pela discriminação dos grupos do que as diferenças na escolaridade. Isto mesmo, já tinha sido comentado na sub-seção 4.2.4.

4.3.2 - Análise das Correlações por Cursos

Adotando um nível de significância de 10%, tem-se que as correlações dentro dos grupos são:

Curso A:

i) Não significativa entre as variáveis contínuas Prin1 e Prin2.

ii) Significativa entre categorias da ocupação do pai, X_5, e participação na vida econômica da família, X_6. As evidências indicam que filhos que trabalham estão associados a ocupação do pai nos estratos inferiores.

iii) Não significativa entre as variáveis contínuas e categóricas.

Curso C:

i) Significativa e positiva entre Prin1 e Prin2, isto indica que no curso C, um bom desempenho geral está associado a melhor desempenho em exatas com relação a humanas (pois Prin2 é um contraste entre as duas áreas).

ii) Não significativa entre as variáveis categóricas.

iii) Não significativa entre as variáveis contínuas e categóricas.

4.3.3 - Aplicação dos Métodos de Discriminação para Misturas

a) Modelo de Posição

- Variáveis consideradas:

  2 contínuas, 1 binária, 1 ternária (substituída por duas binárias).

- Número de categorias multinomiais:

  2*3 = 6

- Número de observações nos grupos

  Curso A: 70

  Curso C: 31

- Probabilidades multinomiais estimadas com um modelo log-linear de segunda ordem (efeitos principais e interação de primeira ordem):

Tabela 4.3.1 - Descrição das categorias dos cursos A e C

X_6	b_1	b_2	$m^{(1)}$	p_{Am} (%)	p_{Cm} (%)	Descrição
0	0	0	1	5,78	9,50	Ocup. pai inf., não trab.
1	0	0	2	17,08	3,41	Ocup. pai inf., trab.
0	1	0	3	19,96	25,98	Ocup. pai sup., não trab.
1	1	0	4	8,61	15,96	Ocup. pai sup., trab.
0	0	1	5	34,29	12,90	Ocup. pai méd., não trab.
1	0	1	6	14,29	32,26	Ocup. pai méd., trab.

(1) Número da categoria $m = 1 + X_6 + 2\,b_1 + 4\,b_2$, pela definiçao 2.1.6

Na tabela anterior nota-se que as maiores probabilidades foram observadas nas categorias 5 e 6, ou seja, para ocupação do pai no estrato médio. No que diz respeito à participação na vida econômica da família, no curso A predominam os alunos que não trabalham e no curso C os alunos que trabalham.

As categorias com menor probabilidade de ocorrência foram 1 e 2, i.e., as categorias nas quais a ocupação do pai pertence ao estrato inferior. Considerando estas duas categorias, os cursos A e C se diferenciam pela participação na vida econômica, no curso A predominam os que trabalham e no curso C os que não trabalham.

Em relação às categorias 3 e 4 nas quais a ocupação do pai pertence ao estrato superior, em ambos os cursos predominam os alunos que não trabalham.

- Análise das médias das variáveis contínuas estimadas com o seguinte modelo de regressão multivariada.

$$\begin{bmatrix} Prin_{1,i} \\ Prin_{2,i} \end{bmatrix} =$$

$$\begin{bmatrix} \nu_1^{(Pr1)} & \alpha_{11}^{(Pr1)} & \alpha_{12}^{(Pr1)} & \alpha_{13}^{(Pr1)} & \beta_{1,12}^{(Pr1)} & \beta_{1,13}^{(Pr1)} \\ \nu_1^{(Pr2)} & \alpha_{11}^{(Pr2)} & \alpha_{12}^{(Pr2)} & \alpha_{13}^{(Pr2)} & \beta_{1,12}^{(Pr2)} & \beta_{1,13}^{(Pr2)} \end{bmatrix} \begin{bmatrix} 1 \\ X_6 \\ b_1 \\ b_2 \\ X_6 b_1 \\ X_6 b_2 \end{bmatrix} + \begin{bmatrix} \varepsilon_1^{(Pr1)} \\ \varepsilon_1^{(Pr2)} \end{bmatrix}$$

$$i = A, C \qquad (4.3.1)$$

onde b_1 e b_2 são duas variáveis binárias pelas quais foi substituída a ternária X_5.

Depois de estimar os parâmetros em (4.3.1) tem-se que os modelos para as médias das variáveis contínuas são:

Curso A:

$$Prin1_A = -1{,}41 + 0{,}11\ X_6 - 0{,}08\ b_1 - 0{,}11\ b_2 + 1{,}35\ X_6 b_1 - 0{,}05\ X_6 b_2 \qquad (4.3.2)$$

$$Prin2_A = -0,41 + 0,66\ X_6 + 0,13\ b_1 + 0,56\ b_2 - 0,27\ X_6b_1 - 1,05\ X_6X_2 \quad (4.3.3)$$

Curso C:

$$Prin1_C = 3,22 + 0,05\ X_6 - 0,26\ b_1 + 0,26\ b_2 - 0,25\ X_6b_1 - 0,49\ X_6b_2 \quad (4.3.4)$$

$$Prin2_C = 0,41 + 0,47\ X_6 - 0,47\ b_1 + 0,09\ b_2 - 0,44\ X_6b_1 - 1,24\ X_6b_2 \quad (4.3.5)$$

No modelo (4.3.2) observa-se que os efeitos principais das variáveis binárias e a interação X_6b_2 não contribuem significativamente à media do desempenho geral do grupo A, $Prin1_A$. Porém, o efeito da interação X_6b_1 é importante pois se $X_6b_1 = 1$, ou seja, se o aluno trabalha e a ocupação do pai pertence ao estrato superior, o valor da média estimada aumentaria consideravelmente, indicando melhor desempenho geral.

No curso C, o modelo (4.3.4) para o desempenho geral, $Prin1_C$, mostra que os efeitos principais e as interações devidas às variáveis categóricas não afetam consideravelmente a média.

Tanto no cuso A como no curso C, os modelos (4.3.3) e (4.3.5) para Prin2 indicam que a média do contraste entre disciplinas de humanas e exatas varia entre as categorias devido ao efeito das variáveis binárias.

Tabela 4.3.2 - Estimativas das médias das variáveis contínuas por categoria e por curso.

X_6	b_1	b_2	Categ (m) [1]	Descrição	Curso A $\overline{Prin1}_A^{(m)}$	Curso A $\overline{Prin2}_A^{(m)}$	Curso C $\overline{Prin1}_C^{(m)}$	Curso C $\overline{Prin2}_C^{(m)}$
0	0	0	1	inf, ñ trab	-1,41	-0,42	3,22	0,41
1	0	0	2	inf, trab	-1,30	0,24	3,17	0,89
0	1	0	3	sup, ñ trab	-1,49	-0,28	2,95	-0,05
1	1	0	4	sup, trab	-0,03	-0,11	2,66	-0,02
0	0	1	5	méd, ñ trab	-1,51	0,15	3,48	0,51
1	0	1	6	méd, trab	-1,45	-0,25	2,94	-0,26
Médias incondicionais das variáveis contínuas					-1,33 [2]	-0,03	3,00	0,04

[1] $m = 1 + X_6 + 2\,b_1 + 4\,b_2$, conforme definiçoes 2.1.6.

[2] Média incondicional de Prin1 no curso A $= \sum_{m=1}^{6} p_{Am}\,\overline{Prin1}^{(m)}$, onde p_{Am}: probabilidade da categoria m no curso A (ver tabela 4.3.1)

A tabela 4.3.2 apresenta as estimativas das médias das variáveis contínuas por categorias e por curso, obtidas substituindo convenientemente as variáveis binárias nos modelos acima.

Analisando esta tabela tem-se que:

. O desempenho geral, Prin1, responde pela principal diferença entre osalunos dos cursos A e C, i.e., a diferença básica entre estes cursos é de média.

. O desempenho geral dos alunos do curso C é notaveimente melhor que o desempenho dos alunos do A em todas as categorias.

. Dentro do curso A tem-se que:

A média de $Prin1_A$ não varia muito entre as categorias 1, 2, 3, 5 e 6, porém é interessante comentar que o pior desempenho geral foi

observado na categoria 5 (ocupação do pai: estrato médio e não trabalha) que tem a maior probabilidade de ocorrência neste curso.

A média de Prin1 na categoria 4 (ocupação do pai: estrato superior, trabalha) se diferencia das outras mostrando que, dentro do curso A, os alunos desta categoria foram os que tiveram melhor desempenho geral (isto era esperado devido ao efeito da interação $X_6 b_1$ no modelo (4.3.2) para a média de $Prin1_A$).

A média do contraste $Prin2_A$ apresenta grande variabilidade. Sinais positivos nas categorias 2 e 5 indicam que, nestas categorias, o desempenho médio nas disciplinas de exatas foi relativamente melhor ao desempenho em humanas (note-se que a categoria 5 é a mais provável neste curso e nela foi observado o pior desempenho geral). Por outro lado, sinais negativos nas outras categorias representam melhor desempenho relativo nas disciplinas de humanas.

. Dentro do curso C nota-se que:

A média do desempenho geral, Prin1, não varia significativamente entre as categorias deste curso.

O contraste $Prin2_C$ é apontado como responsável pela variabilidade entre as categorias de C. Nas três categorias com maior probabilidade de ocorrência: 6 (oc. pai: médio, trabalha), 3 (oc. pai: superior, não tarabalha) e 4 (oc. pai: superior, trabalha), observa-se melhor desempenho na área de humanas relativo as desempenho em exatas. Nas outras categorias: 1, 2 e 5 o desempenho foi melhor em exatas.

- Estimativa da matriz de variância e covariância das variáveis contínuas dadas as categóricas, obtida conforme (3.2.16):

$$\hat{\Sigma} = \begin{bmatrix} 1,33 & -0,07 \\ -0,07 & 0,68 \end{bmatrix} \qquad (4.3.6)$$

- Funções de discriminação calculadas com as estimativas das probabilidades multinomiais na tabela 4.3.1, as estimativas das médias das notas na tabela 4.3.2 e a matriz de dispersão estimada dada em (4.3.6).

Categoria 1: (oc. do pai: inferior, não trabalha)

2,72 - 3,56 Prin1 - 1,60 Prin2

Categoria 2: (oc. do pai: inferior, trabalha)

5,56 - 3,43 Prin1 - 1,31 Prin2

Categoria 3: (oc. do pai: superior, não trabalha)

2,09 - 3,37 Prin1 - 0,69 Prin2

Categoria 4: (oc. do pai: superior, trabalha)

2,03 - 2,02 Prin1 - 0,03 Prin2

Categoria 5: (oc. do pai: médio, não trabalha)

5,01 - 3,80 Prin1 - 0,93 Prin2

Categoria 6: (oc. do pai: médio, trabalha)

1,56 - 3,31 Prin1 - 0,32 Prin2

Chamando $\xi(m)$ à função de discriminação na categoria m tem-se que a regra de classificação nesta categoria é:

Alocar no curso A se $\xi(m) \geq 0$

caso contrário alocar no curso C.

(4.3.7)

Nas funções anteriores nota-se que a variável Prin1 , desempenho geral, é a que mais contribui na discriminação dos cursos. O contraste Prin2, só contribui consideravelmente nas categorias 1 e 2 nas quais a ocupação do pai pertence ao estrato inferior.

- Taxas de erro estimadas pelo método de Lachenbruch.

As proporções estimadas de observações mal classificadas pela regra (4.3.7) são:

Curso A: 7,14%

Curso C: 3,22%

Erro total: 5,19%

b) Função Linear Discriminante de Fisher

- Número de variáveis consideradas: 4
- Regra de alocação baseada na FLD de Fisher

Alocar no curso A se:

$$1{,}80 - 2{,}90\ \text{Prin1} - 0{,}26\ \text{Prin2} + 0{,}35\ X_5 - 0{,}032\ X_6 \geq 0$$

caso contrário alocar no curos C. (4.3.8)

Na função de dicriminação anterior observa-se que a discriminação dos cursos A e C deve-se principalmente ao desempenho geral, Prin1. A contribuição do contraste Prin2 e das variáveis categóricas ocupação do pai e participação na vida econômica (X_5 e X_6), na função discriminante, é pequena comparada com o termo independente e com o efeito de Prin1 que quase definem o valor que assume a função.

- Taxas de erro estimadas pelo método de Lachenbruch, para a regra de classificação (4.3.8)

Curso A: 7,14%

Curso C: 0,00%

Erro total: 3,57%.

c) Funções Lineares Discriminantes Modificadas

Apesar das modificações da FLD de Fisher (sub-seção 3.3.2) terem sido propostas para misturas de variáveis binárias e contínuas, não há maior problema em usar estes métodos para misturas de variáveis contínuas e categóricas em geral, desde que sejam feitas pequenas adaptações similares às discutidas na sub-seção 3.2.3 para o método baseado no modelo de posição. Estas adaptações serão descritas aqui para a variável ternária X_5 que deve ser incluída na análise.

i) Primeira Modificação da FLD

- Variáveis consideradas para a análise:

2 contínuas, 1 binária, 1 ternária que será substituída por duas variáveis binárias (b_1, b_2), conforme definição 2.1.4. Logo, serão consideradas na análise 5 variáveis: 2 contínuas e 3 binárias.

- Transformação do vetor binário em multinomial:

As variáveis categóricas originais geram 2*3 = 6 categorias e somente elas devem ser tomadas em conta na análise.

A transformação que será usada é:

X_6	b_1	b_2	Categ $s^{(*)}$	Z_1	Z_2	Z_3	Z_4	Z_5	Descrição
0	0	0	1	0	0	0	0	0	Ñ Trab, Oc. pai: inf.
1	0	0	2	1	0	0	0	0	Trab, Oc. pai: inf.
0	1	0	3	0	1	0	0	0	Ñ Trab, Oc. pai: sup.
1	1	0	4	0	0	1	0	0	Trab, Oc. pai: sup.
0	0	1	5	0	0	0	1	0	Ñ Trab, Oc. pai: méd.
1	0	1	6	0	0	0	0	1	Trab, Oc. pai: méd.

(*) $s = 1 + X_6 + 2\, b_1 + 4\, b_2$ (1)

Observe-se que o vetor $(X_6, b_1, b_2)'$ não assume nunca valores (1, 1, 1) ou (0, 1, 1) pois $(b_1, b_2) = (1, 1)$ não representa nenhuma das três categorias de X_5.

Depois da transformação, a FLD é então construída nas variáveis Prin1, Prin2, Z_1, ..., Z_5, devendo ser estimados 8 parâmetros para definí-la.

- Função de Discriminação Estimada:

$$2{,}08 - 2{,}97\ \text{Prin1} - 0{,}52\ \text{Prin2} + 2{,}00\ Z_1 - 0{,}23\ Z_2 + 1{,}43\ Z_3 - 1{,}22\ Z_4 - 0{,}90\ Z_5$$

que assume os seguintes valores nas categorias.

(1) Para facilitar a comparaçao das FLD modificadas com o modelo de posiçao as categorias sao identificadas por s = m + 1, e nao por m como na difiniçao 2.1.7.

Categoria 1: Ocup. pai: inferior, não trabalha

$$2{,}083 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Categoria 2: Ocup. pai: inferior, trabalha

$$4{,}083 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Categoria 3: Ocup. pai: superior, não trabalha

$$1{,}853 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Categoria 4: Ocup. pai: superior, trabalha

$$3{,}513 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Categoria 5: Ocup. pai: médio, não trabalha

$$0{,}863 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Categoria 6: Ocup. pai: médio, trabalha

$$1{,}183 - 2{,}97\ Prin1 - 0{,}52\ Prin2$$

Denotando por $\xi(m)$ à função de discriminação na categoria m ($m = 1, 2, \ldots, 6$), tem-se que a regra de classificação nesta categoria é:

Alocar no curso A se $\xi(m) \geq 0$
caso contrário alocar no curso C.

(4.3.9)

As funções discriminantes encontradas representam, por construção, retas paralelas como na figura 4.3.1:

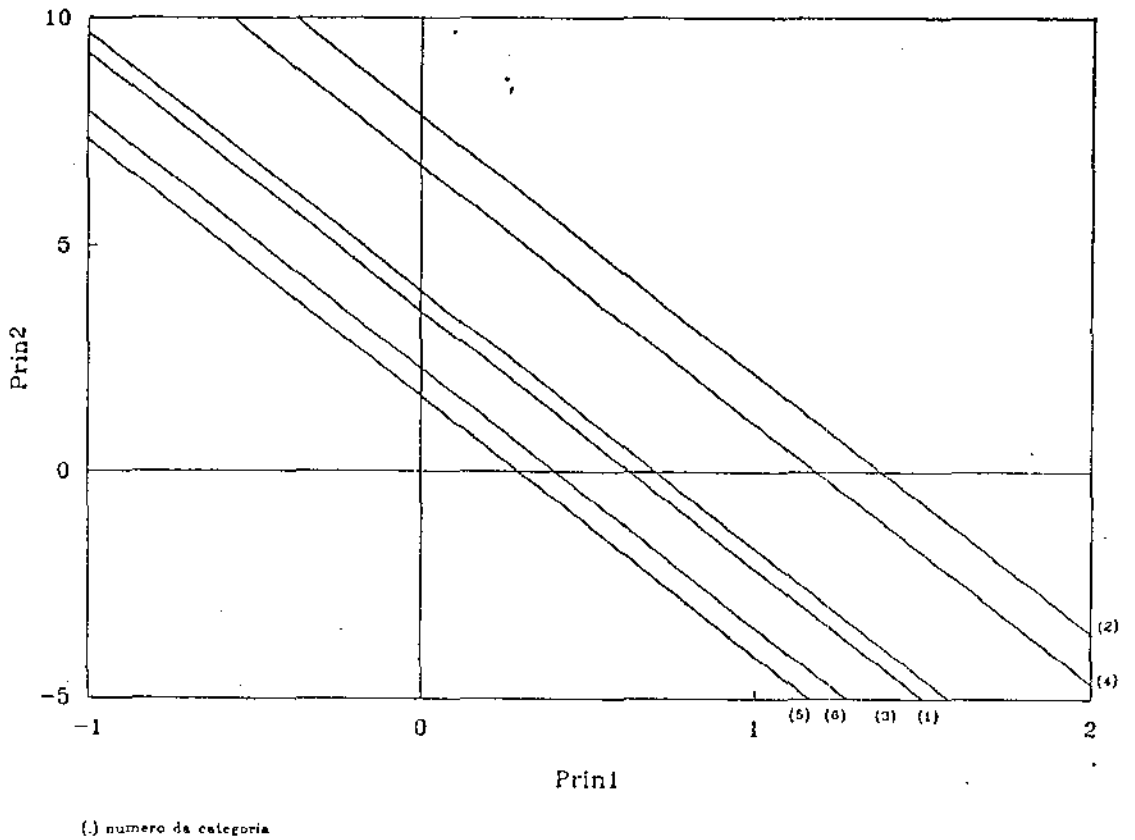

Figura 4.3.1 - Representação das funções discriminantes, baseadas na Primeira Modificação da FLD, para cada uma das 6 categorias em estudo.

Tomando em conta a regra de classificação (4.3.9), os coeficientes negativos de Prin1 e Prin2 nas funções $\xi(m)$ indicam que a pertinência ou classificação do indivíduo no curso A acontece com maior força na medida que o aluno tenha menor desempenho geral e desempenho diferenciado na área de humanas. Isto se repete através de todas as categorias exceto o termo constante (único responsável pelas diferenças entre as discretas).

Na figura 4.3.1 nota-se que a posição das funções no plano Prin1-Prin2 mostra que o melhor desempenho geral foi observado na

categoria 2 (função à direita na figura) e o pior desempenho geral na categoria 5 (função mais à direita da figura), ou seja, melhor desempenho dos ingressantes que trabalham e a ocupação do pai está no estrato inferior, e pior desempenho dos que não tarbalham e a ocupação do pai é de estrato médio.

- Taxas de erro estimadas pelo método de Lachenbruch para a regra de alocação (4.3.9).

  Curso A: 5,71%

  Curso C: 0,00%

  Erro total: 2,86%.

ii) Segunda Modificação da FLD

A segunda modificação proposta em Vlachonikolis e Marriott (1982) consiste em construir a função linear discriminante nas variáveis contínuas Prin1 e Prin2, as variáveis multinomiais Z_1, ..., Z_5 definidas anteriormente e as interações entre elas representadas pelos produtos de forma Z_jPrink para k = 1, 2 e j = 1, ..., 5, logo, o número de parâmetros que devem ser estimados é 2 + 5 + 2*5 + 1 = 18.

- Função de discriminação estimada

$$\begin{aligned} 3,29 &- 4,58\ \text{Prin1} + 1,32\ \text{Prin2} + 1,99\ Z_1 - 1,22\ Z_2 + \\ &+ 1,18\ Z_3 - 1,30\ Z_4 - 1,18\ Z_5 + 2,00\ \text{Prin1}Z_1 - \\ &- 1,14\ \text{Prin2}Z_1 + 1,13\ \text{Prin1}Z_2 - 2,72\ \text{Prin2}Z_2 + 0,91\ \text{Prin1}Z_3 - \\ &- 2,63\ \text{Prin2}Z_3 + 2,06\ \text{Prin1}Z_4 - 2,12\ \text{Prin2}Z_4 + 0,31\ \text{Prin1}Z_5 - \\ &- 1,99\ \text{Prin2}Z_5 \end{aligned} \qquad (4.3.10)$$

a função anterior assume os seguintes valores nas categorias:

Categoria 1: Ocup. pai: inferior, não trabalha

$$3{,}29 - 4{,}58\ \text{Prin1} + 1{,}32\ \text{Prin2}$$

Categoria 2: Ocup. pai: inferior, trabalha

$$5{,}28 - 2{,}58\ \text{Prin1} + 0{,}18\ \text{Prin2}$$

Categoria 3: Ocup. pai: superior, não trabalha

$$2{,}07 - 3{,}45\ \text{Prin1} - 1{,}40\ \text{Prin2}$$

Categoria 4: Ocup. pai: superior, trabalha

$$4{,}47 - 3{,}67\ \text{Prin1} - 1{,}04\ \text{Prin2}$$

Categoria 5: Ocup. pai: médio, não trabalha

$$4{,}59 - 2{,}52\ \text{Prin1} - 0{,}80\ \text{Prin2}$$

Categoria 6: Ocup. pai: médio, trabalha

$$2{,}11 - 4{,}27\ \text{Prin1} - 0{,}67\ \text{Prin2}$$

A regra de classificação para observações da m-ésima categoria ($m = 1, 2, \ldots, 6$), baseada na segunda modificação de FLD é:

$$\text{Alocar no curso A se } \xi(m) \geq 0$$
$$\text{caso contrário alocar no curso C.} \qquad (4.3.11)$$

onde $\xi(m)$ representa a função de discriminação (4.3.10) na categoria m.

Observe-se que neste caso, as retas de classificação já não são paralelas pelo efeito das interações entre variáveis contínuas e multinomiais presentes na função discriminante (4.3.10).

- Taxas de erro estimadas pelo método de Lachenbruch, para a regra de alocação (4.3.11).

Curso A: 4,29%

Curso C: 9,68%

Erro total: 6,98%.

iii) Terceira Modificação da FLD

Nesta modificação já não é necessário transformar as binárias em multinomiais. A função de discriminação é construída considerando as variáveis contínuas, as binárias, os produtos de contínuas e binárias e os produtos cruzados entre binárias.

- Função discriminante estimada:

$$\begin{aligned} 4{,}89 &- 5{,}73\ \mathrm{Prin1} - 1{,}78\ \mathrm{Prin2} - 9{,}98\ b_1 + 1{,}94\ b_2 + \\ &+ 7{,}12\ X_6 - 2{,}45\ b_1X_6 - 3{,}68\ b_2X_6 + 1{,}28\ \mathrm{Prin1}b_1 + \\ &+ 2{,}12\ \mathrm{Prin2}b_2 + 2{,}37\ \mathrm{Prin1}X_6 - 2{,}72\ \mathrm{Prin2}b_1 - 1{,}35\ \mathrm{Prin2}b_2 - \\ &- 1{,}18\ \mathrm{Prin2}X_6 - 1{,}82\ \mathrm{Prin1}b_1X_6 - 4{,}24\ \mathrm{Prin1}b_2X_6 - \\ &- 1{,}23\ \mathrm{Prin2}b_1X_6 - 0{,}49\ \mathrm{Prin2}b_2X_6, \end{aligned} \qquad (4.3.12)$$

que assume os seguintes valores nas categorias:

Categoria 1: Ocup. pai: inferior, não trabalha

$$4{,}89 - 5{,}73\ \mathrm{Prin1} - 1{,}78\ \mathrm{Prin2}$$

Categoria 2: Ocup. pai: inferior, trabalha

$$12{,}01 - 3{,}36\ \mathrm{Prin1} - 2{,}96\ \mathrm{Prin2}$$

Categoria 3: Ocup. pai: superior, não trabalha

$$-5{,}09 - 4{,}45\ \mathrm{Prin1} - 4{,}50\ \mathrm{Prin2}$$

Categoria 4: Ocup. pai: superior, trabalha

$$-0{,}42 - 3{,}90\ Prin1 - 6{,}91\ Prin2$$

Categoria 5: Ocup. pai: médio, não trabalha

$$6{,}83 - 3{,}61\ Prin1 - 3{,}13\ Prin2$$

Categoria 6: Ocup. pai: médio, trabalha

$$10{,}27 - 5{,}48\ Prin1 - 4{,}80\ Prin2$$

Denotando como $\xi(m)$ à função de discriminação (4.3.12) na categoria m (m = 1, 2, ..., 6), a regra de classificação baseada na terceira modificação da FLD é:

Dada uma observação da categoria m,
alocá-la no curso A se $\xi(m) \geq 0$
caso contrário alocar no curso C.

(4.3.13)

Como no caso anterior, as funções de discriminação nas categorias não são paralelas devido à presença de produtos de variáveis contínuas e categóricas na função (4.3.12).

De forma geral, a interpretação da Segunda e Terceira modificação da FLD ganha maior grau de dificuldade pela perda de paralelismo. No entanto, quando há real interesse em mostrar as diferenças nas subpopulações, para entender a classificação através das distintas categorias de forma comparativa, estas funções, que podem não ser paralelas, são melhores. Um caso menos trivial poderia ilustrar melhor a riqueza destes métodos.

- Taxas de erro estimadas pelo método de Lachenbruch.

Curso A: 1,43%

Curso C: 6,45%

Erro total: 3,94%.

Para terminar esta seção, os cinco métodos aplicados serão comparados pela estimativa da taxa de erro total encontrada usando o método de Lanchenbruch (ou *leaving-one-out*) para cada um deles. Estas taxas são:

- Modelo de Posição: 5,19%
- FLD de Fisher: 3,57%
- FLD Modificada 1: 2,86%
- FLD Modificada 2: 6,98%
- FLD Modificada 3: 3,94%,

a melhor classificação resultou da aplicação da primeira modificação da FLD, isto se explica pelo fato de haver uma variável (Prin1), que responde quase totalmente pela discriminação entre os dois grupos e faz com que a diferença entre os cursos, e entre as categorias dentro de cada curso, seja principalmente uma diferença de médias de desempenho geral.

Finalmente, é interessante comentar que as taxas de erro total para o método baseado no modelo de posição e a FLD de Fisher aplicados ao conjunto completo de variáveis (sub-seção 4.2.2) foram 6,08% e 5,90%, respectivamente, portanto, os dois métodos se desempenharam melhor depois da redução de dimensão do problema.

CAPÍTULO 5

CONSIDERAÇÕES FINAIS E CONCLUSÕES

Na prática, é frequente encontrar problemas de discriminação nos quais os grupos ou populações estão caracterizados por observações mistas de variáveis contínuas e categóricas. Do estudo desenvolvido neste trabalho algumas considerações e conclusões são colocadas a seguir no sentido de elevar o patamar de referência ao abordar um problema de discriminação com misturas.

- A principal diferença entre as variáveis contínuas e categóricas no contexto de discriminação é que as últimas geram subpopulações, ou categorias, dentro das quais o padrão de discriminação entre os grupos pode ser diferente.

- Se entre as variáveis categóricas alguma é ordenada, um procedimento possível é tratá-la como se fosse contínua. Krzanowski (1982), afirma que esta seria a melhor solução se o número de categorias é moderado e se há uma boa razão para supor que existe relação linear entre as duas populacões para essa variável. Porém, se a relação não é linear, com certeza, é mais recomendável substituir a variável

ordinal por binárias.

- Em muitos casos, os pesquisadores se limitam a estudar e tentar resolver problemas com misturas de variáveis binárias e contínuas. Isto não é tão restrito como pode parecer, pois qualquer variável categórica pode ser transformada em um vetor de variáveis binárias usando um artifício muito comum na estatística que é a definicão de variáveis *dummy*.

- Num contexto de mistura, a distribuição multinomial aparece como uma ferramenta muito útil pois permite estudar o comportamento das variáveis contínuas dentro de cada uma das categorias geradas pelos distintos arranjos possíveis do vetor de variáveis categóricas.

- O simples exame da média e da matriz de variância-covariância de um vetor aleatório misto, revela com clareza o efeito do tratamento conjunto dos dois tipos de variáveis. Por exemplo, as componentes do vetor de médias das variáveis contínuas são médias ponderadas das médias dentro das subpopulações e a matriz de dispersão das contínuas exibe duas fontes de variabilidade, a primeira dentro das categorias e a segunda entre as categorias. É muito importante que a estrutura particular das observações mistas seja considerada para fazer uma análise adequada dos dados.

- A escolha de um método para discriminar com mistura de variáveis depende, principalmente, da dimensão do problema (i.e., do número de variáveis envolvidas) e dos tamanhos das amostras de treinamento disponíveis, pois o número de parâmetros que devem ser estimados para construir algumas funções de discriminação, cresce excessivamente quando aumenta o número de variáveis consideradas (principalmente quando se inclui muitas variáveis categóricas).

- O comentário anterior sugere como um passo prévio à aplicação de

qualquer método, a redução do número de variáveis que serão consideradas. Para os métodos baseados na FLD de Fisher pode ser usado o procedimento PROC STEPDISC do sistema de programas SAS, para a Discriminação Logística existe uma opção de seleção de variáveis dentro do procedimento PROC LOGISTIC do SAS. Krzanowski (1983) propôs um método de seleção de variáveis categóricas adequado para o método baseado no modelo de posição porém, este não está disponível em forma de programa computacional. No que diz respeito aos métodos kernel e do vizinho mais próximo, não foi encontrado nenhum procedimento de seleção de variáveis apropriado para estes métodos quando usados para discriminação na presença de misturas.

- A aplicação dos métodos de discriminação descritos neste trabalho pode ser feita através dos seguintes programas computacionais:

  . PROC DISCRIM do SAS para a FLD de Fisher e para as Funções Lineares Discriminantes Modificadas. Existem opções deste procedimento que permitem trabalhar ainda com estimação kernel e métodos do vizinho mais próximo.

  . PROC LOGISTIC do SAS para Discriminação Logística.

  . Programa para o método baseado no modelo de posição de W. J. Krzanowski (disponível quando solicitado ao autor).

  . Program ALLOC 80 de J. Hermans, *et al.* para discriminação kernel.

- O método baseado no modelo de posição é o único que pressupõe um modelo para a distribuição dos dados mistos. Sob as condições do modelo este enfoque exibe as mesmas propriedades otimais que a FLD de Fisher aplicada a variáveis explanatórias multinormais e matrizes de variância e covariância comum nas duas populações.

- A maior desvantagem do modelo de posição é o grande número de parâmetros que devem ser estimados, este fato limita o número de variáveis categóricas que podem ser incluídas na análise,

principalmente nos casos em que os tamanhos das amostras de treinamento não são muito grandes.

- O tratamento computacional das Funções Lineares Discriminates Modificadas é totalmente análogo ao da FLD de Fisher, isto é uma grande vantagem pois o *software* necessário para aplicá-las é de fácil acesso.

- A discriminação logística não deve ser considerada como um método de discriminação propriamente dito, senão como um método alternativo para estimar os parâmetros da FLD de Fisher ou suas modificações.

  Na regressão logística os parâmetros são estimados diretamente usando procedimentos iterativos, a diferença com o enfoque usual é que no último os coeficientes são estimados como funções dos estimadores dos vetores de média e matriz de dispersão. Press e Wilson (1978) recomendaram o uso de estimadores de regressão logística em situações onde as suposições de normalidade e/ou de igualdade de matrizes de variância-covariância não são satisfeitas.

- O método de discriminação kernel não faz suposições distribuicionais no entanto, o seu desempenho depende fortemente do parâmetro de alisamento escolhido, logo é indispensável ter um bom conhecimento teórico deste método de estimação de densidades para obter bons resutados na discriminação.

Da literatura cabe destacar as seguintes constatações sobre desempenho comparativo entre os métodos.

- Baseado em um estudo Monte Carlo, Krzanowski (1975) conclui que quando o número de variáveis categóricas aumenta, a taxa de erro do método baseado no modelo de posição tende a crescer ,mas a taxa de erro de FLD permanece estável.

- Em Krzanowski (1977), o desempenho da FLD de Fisher foi comparado com o desempenho do enfoque baseado no modelo de posição, assumindo parâmetros conhecidos. As duas principais conclusões deste trabalho foram:

  a) As maiores taxas de erro associadas à FLD de Fisher foram observadas para o caso de correlações positivas moderadas entre todas as variáveis binárias.

  b) Em geral, a FLD fornece resultados satisfatórios se as funções de discriminação dentro das categorias, baseadas nas variáveis contínuas, são aproximadamente paralelas e têm a mesma orientação. Desvios destas condições implicam em desempenhos cada vez mais pobres da FLD desde que o poder de discriminação individual das funções perde-se quando se pondera sobre todas as categorias. Isto ocorre em situações nas quais a relação entre as médias das variáveis contínuas varia de uma categoria à outra. Agora, correlações altas entre uma variável contínua e uma binária implicam em que a média da variável contínua varie entre as categorias, portanto, uma inspeção preliminar das correlações entre variáveis de diferentes tipos pode mostrar situações em que se deve esperar resultados pobres da FLD. Correlações baixas numa população e altas na outra, ou mudança no sinal de alguma correlação nas populações também foram apontadas como indicações de condições desfavoráveis para a FLD de Fisher.

- Vlachonikolis e Marriott (1982) compararam vários métodos em dois conjuntos de dados. A conclusão principal do trabalho foi que a FLD de Fisher é flexível e eficiente. Contudo, quando há interações presentes, a inclusão de alguns termos apropriados na função contribui na remoção de algumas das mais sérias desvantagens do método.

- Uma das críticas mais frequentes ao método de discriminação logística é o excessivo tempo computacional necessário para

estimação dos parâmetros (ver por exemplo Press e Wilson, 1978). Aitchinson e Aitken (1976) comentaram que quando a discriminação logística é usada, geralmente não é possível estimar as taxas de erro pelo método de Lachenbruch (*leaving-one-out*) pois requer muito tempo computacional.

Para terminar, é interessante comentar que embora a teoria sobre discriminação com misturas esteja bastante desenvolvida, ainda há muitos problemas abertos que devem ser estudados e resolvidos, por exemplo:

- Elaboração de programas computacionais para tratar as observações mistas de maneira adequada, i.e., programas que mostrem o efeito do tratamento conjunto dos dois tipos de variáveis permitindo, por exemplo, recuperar as contribuições da variabilidade dentro e entre as categorias.
- Elaboração de programas computacionais que permitam usar métodos de seleção de variáveis apropriados para misturas como os propostos em Krzanowski (1983), Daudin (1986) ou Krusinska (1989).
- Estudo detalhado das tranformações lineares propostas em Krzanowski (1979) para reduzir o número de variáveis ou simplificar a sua estrutura em problemas com misturas (não necessariamente em discriminação).
- Em geral, estudo de misturas em outros contextos diferentes ao de discriminação, por exemplo, em Análise de Componentes Principais ou Análise de Conglomerados.
- Aplicações de outras medidas de distância entre observações mistas para o método de discriminação segundo o vizinho mais próximo.

APÊNDICE A

Conjunto de dados usado na aplicação do capítulo 4.

Curso A

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
5.88	2.88	5.38	4.75	3.63	6.38	4.88	2.63	0	0	1	0	2	0
3.75	2.25	1.75	4.50	1.63	3.63	4.25	5.88	1	1	1	0	1	0
3.88	3.13	4.50	4.25	1.63	4.88	3.00	5.25	1	1	1	0	2	0
6.13	1.63	6.38	5.38	2.75	4.88	4.75	8.75	1	1	1	0	1	1
2.13	1.38	1.00	2.38	0.38	4.38	3.25	3.00	0	0	1	1	2	0
4.13	2.25	6.63	3.00	2.00	3.50	3.63	4.25	1	0	0	0	1	1
2.00	0.50	0.63	2.88	0.63	4.25	3.25	4.88	0	0	0	0	2	0
2.00	2.63	3.63	1.88	0.50	2.13	3.63	3.00	1	0	1	0	1	0
3.00	5.00	4.50	5.63	1.88	5.38	3.38	5.00	1	1	1	0	1	0
2.63	2.50	2.13	2.75	1.38	3.38	3.63	2.88	0	1	0	1	3	1
3.25	2.13	4.75	5.13	3.50	4.50	3.00	3.13	0	0	1	0	3	0
6.00	4.25	3.88	4.75	3.50	5.75	3.25	4.25	1	1	1	0	2	0
3.88	3.38	2.13	5.38	2.38	4.75	5.88	5.13	0	1	1	0	1	0
2.63	2.13	1.00	4.88	0.88	5.63	4.13	5.88	1	0	0	0	1	0
4.38	2.25	2.13	3.63	1.50	4.00	3.50	4.88	1	0	1	0	1	0
6.00	5.50	7.75	5.50	6.13	5.50	5.50	9.38	1	1	1	0	1	1
3.25	1.25	2.25	5.00	4.88	6.13	4.00	3.88	0	1	0	1	2	0
5.13	3.00	4.38	4.38	3.00	5.00	4.88	6.75	1	1	1	0	1	0
1.38	1.00	1.75	3.50	1.25	3.13	3.00	2.13	0	0	0	1	2	0
2.38	2.38	5.38	2.88	2.25	3.50	5.00	3.50	0	0	1	0	2	0
3.38	1.88	5.25	3.38	1.88	4.00	3.88	3.75	1	1	1	0	2	0
4.13	3.00	4.75	4.00	2.63	4.38	3.38	4.38	0	0	1	0	3	0
4.13	1.88	2.50	4.25	4.38	4.13	3.75	3.38	0	1	0	1	3	0

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
4.38	1.25	0.50	5.25	1.63	6.88	3.38	6.25	0	1	1	1	3	1
3.38	2.00	2.63	3.25	2.88	3.25	4.38	4.00	0	0	0	1	3	1
3.63	2.13	6.00	5.13	2.75	4.38	3.00	4.38	0	0	0	0	2	1
4.75	3.50	4.50	6.00	4.88	5.38	5.13	4.50	1	1	1	0	2	0
6.25	2.75	6.25	5.63	7.63	4.63	6.75	6.38	1	1	1	0	2	0
3.75	3.13	3.00	4.63	2.38	3.63	3.88	1.38	1	1	1	0	2	0
2.88	3.00	2.50	3.13	2.25	4.00	3.25	3.88	1	1	1	0	2	0
3.88	3.50	3.13	3.38	1.00	3.63	3.50	1.88	0	0	1	0	2	1
3.50	2.00	3.25	3.88	2.75	4.13	4.88	6.50	1	0	0	0	2	1
4.00	1.13	0.25	4.75	1.38	5.25	4.38	1.25	0	0	0	0	2	1
4.00	3.38	4.63	4.88	2.00	4.63	4.25	4.13	0	1	1	0	3	1
5.63	3.88	5.13	4.63	3.50	4.50	3.13	7.63	0	1	1	0	1	0
4.00	1.63	2.38	3.75	1.88	4.50	3.50	4.50	1	1	0	0	2	0
3.25	0.75	1.88	1.88	0.75	2.25	3.38	0.75	0	1	0	1	1	0
3.50	2.13	5.50	2.25	6.13	3.75	4.75	6.75	0	0	1	1	2	0
4.50	3.25	6.00	3.63	4.25	5.13	5.13	3.63	1	1	1	0	3	1
4.75	2.25	3.38	5.63	0.75	5.38	3.38	6.50	0	0	1	0	2	0
3.88	2.00	2.88	2.75	1.50	2.88	3.25	3.25	0	0	0	0	2	0
3.25	2.13	5.88	3.25	3.13	4.88	3.75	4.25	1	0	0	0	2	1
4.25	1.63	3.38	4.13	2.63	4.25	3.63	7.38	1	1	1	0	1	0
2.50	3.13	4.38	4.25	1.75	4.88	3.88	3.63	0	0	1	1	3	1
2.00	1.63	2.75	2.50	0.50	2.63	3.13	1.63	0	1	0	1	2	0
2.75	3.25	3.63	4.50	4.00	4.13	3.25	5.75	0	0	0	0	2	0
4.38	2.00	4.75	5.13	2.88	4.25	3.50	5.63	1	1	1	0	1	0
3.88	1.75	3.13	5.75	2.75	3.25	3.50	5.00	0	0	0	1	2	1
4.00	1.13	3.50	2.63	3.50	2.63	4.75	4.88	0	1	1	0	2	0
6.13	1.63	4.50	5.50	1.38	4.50	3.38	6.75	0	1	1	0	1	0
1.88	1.25	4.38	4.50	3.25	4.00	6.00	7.88	1	1	1	0	2	0
2.63	2.25	5.25	4.00	2.75	4.13	5.00	4.13	0	1	1	0	1	1
2.13	3.63	3.88	4.50	2.75	3.88	4.13	2.00	0	0	1	0	3	1
3.00	3.25	5.88	4.38	7.00	4.75	6.88	7.13	0	1	0	1	3	1

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
2.75	2.88	3.25	4.88	2.13	6.25	3.38	3.50	0	0	0	0	2	1
5.50	2.38	5.00	3.13	3.75	3.13	5.88	3.25	0	1	1	0	2	0
2.88	1.50	2.25	1.38	1.25	2.00	3.13	1.75	0	1	1	0	2	0
2.63	1.75	3.38	3.25	1.38	3.13	5.38	1.75	0	0	0	0	2	1
2.13	1.75	3.25	3.50	0.75	4.25	3.75	2.75	0	0	1	0	1	1
2.38	0.88	1.63	2.38	1.50	2.88	3.00	0.50	0	1	1	0	1	0
4.25	3.38	3.63	5.75	2.63	4.00	3.50	5.38	0	1	0	0	2	1
3.63	1.50	2.63	4.63	1.63	5.00	3.13	4.25	0	0	1	0	3	1
4.63	1.63	3.13	4.63	1.13	4.63	3.50	6.00	0	0	1	0	2	1
3.75	2.00	4.75	3.13	1.25	3.50	4.25	7.13	1	1	1	0	2	0
5.38	1.75	2.25	4.25	0.88	4.75	3.13	2.25	0	0	1	0	3	0
3.13	3.63	3.75	3.63	1.75	3.00	3.75	5.25	1	1	1	0	1	0
5.88	2.75	7.25	5.00	2.50	5.38	5.75	9.13	1	0	0	0	1	1
2.88	2.00	2.00	2.50	0.75	3.00	4.25	2.25	1	1	0	1	3	1
4.13	2.88	4.75	5.25	5.25	4.25	4.88	0.25	0	0	0	0	3	1
3.63	2.88	3.38	3.75	4.63	3.25	3.50	4.50	0	0	1	1	3	1

Curso B

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
6.88	6.63	8.38	6.38	8.75	5.75	7.63	6.88	1	1	1	0	2	0
6.88	6.25	7.63	7.38	4.38	7.38	3.75	7.25	1	1	1	0	1	0
6.13	5.63	6.38	6.75	6.63	5.38	5.88	5.63	1	1	1	0	2	0
5.50	6.25	8.13	5.75	7.75	6.25	5.25	9.25	1	1	0	0	2	1
5.13	6.88	8.38	7.00	8.13	6.88	7.38	8.13	0	0	1	0	1	0
4.38	7.00	8.25	6.00	9.63	5.50	9.00	7.63	1	1	1	0	2	0
5.75	6.50	7.38	5.13	6.00	6.25	5.75	8.25	0	0	1	0	1	0
7.88	8.63	7.75	8.13	8.88	6.25	9.38	9.38	1	1	1	0	1	0
5.88	4.63	5.38	6.63	7.75	5.63	6.75	9.50	1	1	1	0	2	0

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
4.63	6.88	8.75	7.50	8.63	6.50	7.25	8.75	1	1	1	0	2	0
6.25	5.50	7.75	7.13	7.75	6.38	8.63	10.00	1	1	1	0	1	0
4.75	6.50	8.38	7.50	9.50	6.38	9.38	8.63	0	0	0	0	3	0
6.63	6.25	7.38	7.50	7.38	6.00	5.88	5.50	1	1	1	0	2	0
6.25	6.50	8.13	6.50	9.50	6.00	9.25	9.75	1	1	1	0	2	0
5.63	4.75	7.38	6.63	7.88	6.50	6.50	8.75	0	0	0	0	2	0
5.13	4.50	6.63	5.88	7.75	5.75	8.75	8.25	1	0	0	0	3	1
6.75	7.88	8.75	7.38	7.00	5.50	8.00	9.75	1	1	1	0	1	0
4.75	3.50	4.75	6.50	8.63	6.63	7.50	9.00	1	1	1	0	1	0
5.00	4.75	7.13	6.25	7.63	6.38	8.00	9.38	0	1	1	0	2	0
6.75	4.88	7.13	6.38	7.13	5.63	8.63	8.00	0	0	0	0	2	0
6.50	4.13	6.63	7.63	5.63	5.88	6.88	9.25	1	1	1	0	1	0
5.25	6.63	8.38	7.63	7.63	6.25	6.25	9.25	1	1	1	0	1	0
6.50	6.13	9.50	8.63	9.25	6.50	10.00	9.88	1	1	1	0	1	0
5.63	5.13	7.25	6.88	6.13	6.75	6.38	8.00	1	1	1	0	2	0
8.00	8.63	7.88	8.13	9.13	7.75	6.25	9.63	1	1	1	0	3	0
7.25	6.25	8.13	6.75	4.25	6.00	7.25	9.25	1	1	1	0	1	0
4.75	5.38	7.75	6.38	7.13	6.25	7.13	7.50	1	1	1	0	2	0
4.88	4.88	6.13	6.50	5.00	5.75	5.50	7.38	0	1	1	0	2	0
6.00	5.63	7.50	6.00	6.00	6.88	7.38	8.88	1	1	1	0	3	0
6.38	4.75	6.63	6.38	7.00	6.88	6.50	9.25	1	0	1	0	1	0
6.13	6.75	8.25	6.38	7.88	6.00	7.50	8.13	1	1	1	0	1	0
5.25	6.88	6.13	7.25	5.00	7.38	3.88	9.38	0	0	1	0	1	0
6.13	5.13	6.50	6.13	7.88	4.38	7.50	8.38	1	1	1	0	2	1
5.25	7.25	9.00	6.75	8.63	6.25	9.25	8.63	0	1	1	0	1	0
5.00	4.38	7.38	6.75	7.50	5.63	7.38	7.63	1	1	1	0	1	0
5.38	5.50	7.63	6.13	7.38	5.13	7.75	8.88	1	1	1	0	1	0
6.50	7.50	8.75	6.63	7.50	6.50	5.75	7.88	1	1	1	0	1	0
4.50	4.88	6.50	6.88	6.63	6.38	7.13	6.50	1	1	1	0	2	0
6.00	5.75	7.75	7.25	8.13	6.63	5.63	5.63	1	1	1	0	2	0
7.38	4.75	8.38	8.13	6.13	5.63	6.38	9.13	1	1	1	0	1	0

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
6.50	5.00	8.25	6.50	9.63	6.25	9.00	9.63	0	1	1	0	1	0
5.88	5.88	7.25	6.75	8.88	5.13	8.88	9.88	1	1	1	0	1	0
6.88	4.75	7.38	6.38	7.88	6.00	7.25	9.38	1	1	1	0	2	0
5.38	6.38	7.25	7.38	8.25	6.38	6.00	9.13	1	1	1	0	1	0
4.63	6.88	7.88	6.50	7.63	5.63	7.88	8.63	1	1	1	0	1	0
5.13	4.25	7.75	6.38	8.25	5.63	6.75	9.13	1	0	1	1	1	1
4.50	4.75	7.75	6.50	8.25	6.50	6.75	8.38	0	0	1	0	2	0
6.38	5.00	6.38	7.25	2.50	7.25	6.13	9.88	0	0	0	0	2	1
6.50	4.13	8.25	6.63	7.00	4.88	6.38	9.25	1	1	1	0	2	0
6.38	5.38	7.75	7.13	9.25	6.13	7.38	9.88	1	1	1	0	1	1
6.63	8.38	8.38	6.25	9.13	7.25	8.25	7.63	0	1	1	0	1	0
6.50	7.63	8.50	5.75	8.50	5.38	6.75	9.75	1	1	1	0	2	0
5.88	6.38	7.25	6.25	8.13	5.50	7.75	9.25	1	0	0	0	1	1
5.50	4.75	7.75	6.75	6.63	5.25	8.00	8.25	1	1	1	0	1	0
5.13	5.88	7.88	6.50	7.88	5.63	8.75	7.63	1	1	1	0	1	0
6.75	7.50	8.25	8.38	8.63	6.88	9.13	9.38	1	1	1	0	2	0
6.00	5.25	6.50	7.75	5.50	6.38	6.25	8.38	0	0	0	0	2	1
7.00	4.25	5.25	7.63	4.63	6.75	5.50	8.63	1	1	1	0	2	1
6.38	5.50	5.00	6.13	7.38	6.50	7.38	7.00	1	1	1	0	2	0
6.00	6.38	8.63	7.13	6.88	6.88	5.13	9.13	1	1	1	0	1	0
3.75	5.75	7.00	7.50	8.13	6.50	8.63	8.50	1	1	1	0	2	0
4.25	3.38	8.13	6.50	7.38	6.00	8.25	9.00	0	1	0	0	3	0
6.75	5.88	8.75	6.75	7.25	6.38	7.38	9.38	1	1	0	0	1	0
5.25	5.88	7.50	5.88	7.25	5.75	7.50	5.25	0	1	1	0	2	0
6.88	5.75	6.63	7.13	6.25	6.25	7.00	5.13	0	0	1	0	3	0
5.75	4.38	8.00	7.25	7.00	7.13	8.25	7.00	1	0	0	0	1	0
5.38	6.25	8.00	6.88	6.63	7.63	5.00	7.13	1	1	1	0	1	0
6.63	5.00	8.13	7.00	6.75	6.75	7.25	9.63	1	1	1	0	1	0
7.25	5.50	5.88	6.75	6.63	5.50	5.13	8.38	1	1	1	0	1	1
5.63	6.50	7.88	7.13	7.75	5.63	6.63	6.88	0	1	1	0	1	0
4.88	7.00	6.75	6.75	6.63	6.63	4.50	9.63	1	1	1	0	1	0

port	biol	quim	his	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
6.00	5.88	7.88	7.50	9.25	5.38	7.75	8.38	1	1	1	0	1	0
6.38	3.63	7.38	7.75	6.50	5.88	8.63	9.88	1	1	1	0	1	0
7.13	5.38	8.75	5.88	9.13	5.63	8.38	6.63	1	1	0	1	2	0
6.75	5.00	8.63	7.50	4.00	7.13	5.25	9.50	1	0	0	0	1	0
8.13	5.00	8.38	6.88	8.50	5.88	9.25	9.13	1	1	1	0	1	0
5.13	4.38	6.63	6.25	8.50	6.13	9.00	8.38	0	0	1	0	1	0
7.00	5.00	7.88	6.25	8.25	5.88	7.13	9.38	1	1	1	0	2	0
6.00	6.38	5.75	6.75	6.50	6.25	5.88	8.75	0	0	1	0	3	0
6.50	6.13	5.88	7.50	8.38	5.75	7.88	6.25	0	0	1	0	2	0
5.63	6.13	7.25	6.13	6.13	5.88	8.13	6.88	1	1	1	0	3	0
4.88	6.38	6.50	6.63	5.75	6.00	7.38	8.75	1	1	1	0	2	0
6.88	6.13	8.13	7.88	6.38	6.75	7.25	8.50	1	1	1	0	1	0
5.00	6.00	8.00	6.88	7.13	5.38	7.50	6.50	1	1	0	0	1	0
5.50	7.63	7.88	7.50	9.25	6.63	8.00	9.00	1	1	1	0	1	0
5.38	7.25	8.13	7.50	7.50	5.63	7.75	4.75	1	0	0	0	1	0
7.13	7.63	7.38	5.88	6.25	7.25	7.25	7.25	1	1	1	0	1	0
6.38	5.25	5.50	7.13	6.63	6.50	6.38	9.75	1	1	1	0	1	0
5.63	5.00	7.25	7.63	6.88	5.75	8.25	6.63	0	0	0	0	1	0
5.38	6.38	8.25	7.00	8.75	6.38	7.13	8.50	0	0	0	0	2	0

Curso C

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
4.75	7.13	7.75	7.13	6.25	6.50	5.88	7.50	0	0	1	0	1	0
5.38	5.50	6.38	5.13	6.13	4.63	8.38	6.13	0	0	0	0	1	1
8.13	6.38	6.63	7.00	6.88	6.88	5.00	6.88	0	0	0	1	1	1
5.25	6.00	7.75	6.00	8.50	5.88	6.13	6.88	1	1	1	0	2	0
6.13	4.88	5.38	6.88	5.88	6.13	6.25	6.50	0	0	1	0	3	0
4.75	6.63	8.00	7.63	8.25	6.88	5.25	8.25	1	1	1	0	2	0

port	biol	quim	hist	fis	geog	mat	lest	x1	x2	x3	x4	x5	x6
6.25	5.63	7.38	5.75	7.00	6.63	7.50	5.63	0	0	1	0	2	0
5.75	6.75	6.38	6.75	6.00	6.00	5.75	9.38	0	1	0	0	2	1
5.13	6.13	7.25	6.25	8.13	6.00	4.25	8.88	1	1	1	0	1	0
3.63	5.25	7.38	6.00	5.63	6.75	5.25	7.38	1	1	1	0	1	0
7.50	5.88	6.00	7.50	5.13	7.63	7.13	9.88	1	0	0	0	2	1
6.75	7.63	7.38	7.50	6.13	6.63	6.00	8.50	1	1	1	0	2	1
5.75	5.00	8.38	6.50	6.50	5.63	7.75	7.63	0	1	1	0	2	1
7.75	3.63	5.25	5.88	4.25	6.88	4.63	8.13	0	1	0	0	2	1
5.13	4.00	6.13	6.88	6.38	5.25	6.50	9.75	0	1	1	0	2	1
6.00	3.63	6.13	6.25	4.75	6.00	7.38	6.63	0	0	0	0	2	1
5.75	5.38	7.50	7.63	5.50	6.38	6.25	9.50	1	1	1	0	3	0
5.25	4.00	7.63	6.50	4.50	5.38	4.63	9.38	1	1	1	0	1	0
5.75	7.63	8.63	7.63	9.38	6.38	9.13	6.50	1	1	1	0	1	0
4.25	4.88	8.50	5.38	5.88	7.25	5.00	9.88	0	0	0	0	1	1
6.13	1.88	4.50	5.88	4.88	7.38	6.13	9.88	1	0	0	0	1	1
5.50	5.38	7.13	7.63	5.00	7.00	4.00	8.38	0	0	1	1	2	1
5.75	4.13	6.38	6.50	6.75	5.88	8.13	8.50	0	0	0	0	3	1
7.00	4.75	7.00	6.13	4.75	6.13	6.00	5.50	0	0	1	0	2	1
5.38	4.88	7.88	6.63	5.75	4.75	5.00	8.25	1	1	1	0	1	1
6.75	4.63	7.25	7.00	4.13	6.00	4.13	9.50	1	1	1	0	1	0
5.38	6.13	6.63	6.38	3.88	6.25	5.25	8.88	0	0	1	0	1	0
3.88	3.00	5.88	6.88	7.50	6.25	5.25	7.00	0	1	0	0	2	1
5.38	7.00	8.00	5.88	6.50	5.25	7.88	8.50	0	0	0	1	3	0
6.00	5.63	7.50	7.38	7.25	6.75	5.75	7.50	0	1	1	0	2	0
5.50	6.00	6.13	7.25	5.88	6.25	5.00	7.50	0	1	1	0	1	0

ANALISE PRELIMINAR DOS DADOS

Neste apêndice apresentam-se análises iniciais dos dados acima por curso e por tipo de variáveis. As variáveis contínuas são caracterizadas pela média e desvio padrão amostrais enquanto que o número de observações nas categorias das variáveis categóricas é apresentado na forma de tabela de contigência. Por último comenta-se as correlações significativas entre as variáveis em estudo para cada curso.

Média e desvio padrão das variáveis contínuas

Tabela A.1 - Estimativas da média e desvio padrão das notas por cursos.

	Curso A		Curso B		Curso C	
Matéria	Média	D.P.	Média	D.P.	Média	D.P.
Port	3,74	1,19	5,94	0,90	5,73	1,02
Biol	2,37	0,96	5,81	1,14	5,33	1,34
Quim	3,71	1,65	7,50	0,96	6,67	1,00
Hist	4,07	1,11	6,86	0,66	6,64	0,70
Fis	2,52	1,57	7,38	1,38	6,11	1,34
Geog	4,23	1,03	6,18	0,63	6,24	0,72
Mat	4,04	0,94	7,24	1,30	6,02	1,33
L_est	4,42	2,07	8,37	1,26	8,02	1,30

Na tabela A.1 observa-se que, considerando as médias das notas nas matérias avaliadas, o desempenho dos alunos dos cursos B e C é similar, sendo que as médias correspondentes ao curso B são

ligeiramente maiores. Em todos os casos, as médias das notas no curso A são consideravelmente menores que nos outros cursos e ainda ficam abaixo da nota média aprovatória, ou seja são menores que cinco.

Com relação à variabilidade das notas observa-se que os valores dos desvios nos cursos B e C são próximos sendo que em geral, são um pouco menores para B. Nos dois casos a variabilidade não é grande. Por outro lado, a maior variabilidade é observada nas notas dos alunos do curso A.

Um fato interessante de ser comentado é que nos três cursos a menor média foi observada em biologia e a maior média em língua estrangeira.

- Tabelas de Contingência para variáveis categóricas

Tabela A.2 - X_1: Tipo de estabelecimento de ensino do primeiro grau.

	Curso A	Curso B	Curso C	Total
Pred. Estab. Público (0)	43 (61,43)	23 (25,56)	19 (61,29)	85
Pred. Estab. Partic. (1)	27 (38,57)	67 (74,44)	12 (38,71)	106
Total	70	90	31	191

(.) Percentagem na coluna

Nos cursos A e C observa-se que aproximadamente 60% dos alunos cursaram o primeiro grau predominantemente em estabelecimentos públicos. Para o curso B este padrão se inverte e adquire maior força,

74,4% dos alunos estudou primeiro grau predominantemente em escolas particulares e somente 25,6% estudou em escolas públicas.

Tabela A.3 - X_2: Tipo de estabelecimento de ensino do segundo grau.

	Curso A	Curso B	Curso C	Total
Pred. Estab. Público (0)	33 (47,14)	22 (24,44)	14 (45,16)	69
Pred. Estab. Partic. (1)	37 (52,86)	68 (75,56)	17 (54,84)	122
Total	70	90	31	191

(.) Percentagem da coluna

A tabela A.3 mostra que nos três cursos a maior parte dos alunos fez o segundo grau predominantemente em estabelecimentos particulares. No curso B a maioria é clara com uma proporção de 75,56% para $X_2 = 1$ porém, nos cursos A e C a diferença nas proporções das categorias de X_2 não é tão acentuada (é menor que dez porcento).

Tabela A.4 - X_3: Curso de segundo grau concluído.

	Curso A	Curso B	Curso C	Total
Técnico (0)	25 (35,71)	17 (18,89)	11 (35,48)	53
Comum (1)	45 (64,29)	73 (81,11)	17 (64,52)	138
Total	70	90	31	191

(.) Percentagem da coluna

Na tabela anterior observa-se que nos três cursos, a proporção de alunos que terminaram o segundo grau comum é consideravelmente maior que a proporção de alunos que terminaram em escolas técnicas.

Em particular, a diferença de proporções entre as categorias 0 e 1 de X_3 (curso de segundo grau concluído) para o curso B é mais acentuada que nos cursos A e C.

Tabela A.5 – X_4: Período em que cursou o segundo grau.

	Curso A	Curso B	Curso C	Total
Pred. Diurno (0)	55 (78,57)	88 (97,78)	28 (90,32)	171
Pred. Noturno (1)	15 (21,43)	2 (2,22)	3 (9,68)	20
Total	70	90	31	191

(.) Percentagem da coluna

A proporção de alunos que fez o segundo grau predominantemente em período noturno é muito baixa nos curos B e C (menos de 10%). No curso A esta proporção aumenta até 21,43%, mas é ainda baixa comparada com o 78,6% dos alunos deste curso que estudaram predominantemente no período diurno.

Tabela A.6 - X_5: Categoria à qual pertence a ocupação do pai.

Estrato	Curso A	Curso B	Curso C	Total
Superior (1)	20 (28,57)	49 (54,44)	13 (41,94)	82
Médio (2)	34 (48,57)	33 (36,67)	14 (45,16)	81
Inferior (3)	16 (22,86)	8 (8,89)	4 (12,90)	28
Total	70	90	31	191

(.) Percentagem da coluna

Considerando a ocupação do pai nota-se que nos cursos B e C há predominância dos estratos médios e superior. No curso A, predomina o estrato médio; as proporções observadas nos outros dois estratos, superior e inferior, foram de menos de trinta porcento.

Tabela A.7 - X_6: Participação na vida econômica da família.

	Curso A	Curso B	Curso C	Total
Não Trab (0)	42 (60,00)	80 (88,89)	15 (48,39)	137
Trabalha (1)	28 (40,00)	10 (11,11)	16 (51,61)	54
Total	70	90	31	191

(.) Percentagem da coluna

A tabela A.7 mostra que as proporções correspondentes aos alunos que trabalham ou não trabalham foram diferentes nos três cursos: no curso A, 40% dos ingressantes exercem atividades remuneradas, no curso B esta proporção é bem menor: 11,1%, já no curso C a percentagem de alunos que trabalha é de 51,6%, só um pouco maior que a percentagem

dos que não trabalham, 48,4%.

As análises univariadas das variáveis contínuas e das variáveis categóricas feitas acima contribuem à formação de alguma idéias sobre as possíveis fontes de discriminação entre os cursos quando considerados dois a dois. Estas idéias são expostas a seguir:

Curso A e Curso B: As médias das notas são consideravelmente maiores em B logo, espera-se que estas variáveis tenham contribuição significante à discriminação. Em relação às variáveis categóricas, as tabelas de contingência mostraram importantes diferenças nas proporções observadas para os cursos A e B nas distintas categorias portanto, estas variáveis também devem contribuir na discriminação destes cursos.

Curso A e curso C: Como no caso anterior, as médias das notas obtidos pelos alunos do curso A são menores que as do curso C em todas as matérias avaliadas portanto, espera-se que as notas discriminem estes cursos. O perfil dos alunos dos cursos A e C através das variáveis X_1, X_2, X_3 e X_4, referentes à escolaridade são semelhantes logo, espera-se que estas variáveis não contribuam na separação dos cursos. A diferença nas proporções observadas nestes cursos, é maior para as categorias das variáveis X_5 e X_6, referentes ao aspecto econômico, isto as aponta como possíveis fontes de discriminação.

Curso B e curso C: Considerando a proximidade das médias das notas nestes cursos, não é intuitivo esperar que estas variáveis tenham contribuição significante na discriminação. por outro lado, as diferenças nas proporções observadas em alguma tabelas de contingência acima, levam a pensar que os alunos dos cursos curso B e C se diferenciam mais pelas características sócio-econômicas do que pelo seu desenpenho acadêmico.

- Correlações dentro de cada curso

A análise das correlações será feita aqui por cursos e por tipo de variável, adotando como ponto de corte um nível de significância de 10%[1]. O coeficiente de correlção utilizado é o usual, de Pearson, para todas as variáveis consideradas.

i) Curso A

. Correlacões entre variáveis contínuas

Tabela A.8 - Correlação entre as notas no curso A escolhidas para exame.

	Matérias						
	Port	Biol	Quim	Hist	Fis	Geog	Mat
Biol	0,29 (0.01)						
Quim	0.43 (0.00)	0.49 (0.00)					
Hist	0.50 (0.00)	0.39 0.00)	0.26 (0.03)				
Fis	0.37 (0.00)	0.38 (0.00)	0.56 (0.00)	0.36 (0.00)			
Geog	0.41 (0.00)	0.28 (0.02)	— —	0.70 (0.00)	0.25 (0.03)		
Mat	0.24 (0.05)	0.20 (0.09)	0.45 (0.00)	— —	0.58 (0.00)	0.13 (0.28)	
L_est	0.44 (0.00)	0.22 (0.06)	0.43 (0.00)	0.43 (0.00)	0.33 (0.01)	0.37 (0.00)	0.32 (0.01)

(.) Prob > |R|

(1) A hipótese testada é H_0: $\rho = 0$, onde ρ representa o coeficiente de correlaçao de Pearson.

A tabela anterior mostra que existe evidência de associação linear no desempenho dos alunos do curso A, nas distintas disciplinas avaliadas (exceto Mat e Geog, Mat e Hist, Geog e Quim), isto indica que se deve esperar alunos com desempenho geral homogêneo (seja bom ou não).

. Correlações entre as variáveis categóricas

Tabela A.9 - Correlações entre as variáveis categóricas no curso A escolhidas para exame.

	Variáveis Categóricas				
Var. Categ.	X_1	X_2	X_3	X_4	X_5
X_2	0,34 (0,00)				
X_3	— —	0,25 (0,04)			
X_4	-0,34 (0,00)	— —	-0,34 (0,00)		
X_5	-0,39 (0,00)	-0,20 (0,10)	— —	0,38 (0,00)	
X_6	— —	-0,28 (0,02)	-0,30 (0,01)	— —	0,31 (0,01)

(.) Prob > |R|

Da tabela anterior tem-se que:

. A associação entre X_1 e X_2 indica que a tendência foi manter o tipo de estabelecimento de ensino ao passar do primeiro ao segundo grau.

. A relação de X_1 e X_2 com X_5 mostra que o tipo de estabelecimento de

ensino, tanto no primero como no segundo grau, apresentou associação significativa com a ocupação do pai, isto indica que pais com ocupações nos estratos inferiores tenderam a matricular os filhos em escolas públicas, no entanto, pais com ocupações consideradas em estratos superiores escolheram escolas particulares.

. A participação na vida econômica, X_6, também está relacionada à ocupação do pai, X_5. A evidência indica associação entre alunos que trabalham e pais com ocupações nos estratos inferiores.

. A associação entre X_6 e X_2 e entre X_6 e X_3 (participação na vida econômica da família com curso e tipo de escola de segundo grau) mostra uma tendência dos alunos de escolas públicas e técnicas a trabalhar.

. No que diz respeito a X_4, período de estudo no segundo grau este se relaciona significativamente com X_1, X_3 e X_5, ou seja com o tipo de estabelecimento de ensino de primeiro grau, curso feito no segundo grau e ocupação de pai, respectivamente. Assim, segundo grau em período diurno está associado a primeiro grau em escolas particulares, cursos comuns e ocupação do pai nos estratos superiores. Por outro lado, período noturno se associa a primeiro grau público, curso técnico no segundo grau e ocupação do pai nos estratos baixos.

. **Correlação entre variáveis contínuas e categóricas**

Tabela A.10 - Correlação entre variáveis contínuas e categóricas no curso A escolhidas para exame.

	Variáveis Categóricas				
Matéria	X_1	X_2	X_3	X_4	X_5
Port	0,21	0,22	0,26	-0,30	—
	(0,08)	(0,06)	(0,03)	(0,01)	—
Biol	0,26	—	0,27	-0,25	—
	(0,03)	—	(0,03)	(0,04)	—
Quim	0,28	—	0,20	-0,30	—
	(0,02)	—	(0,09)	(0,01)	—
Hist	—	—	—	-0,23	—
	—	—	—	(0,06)	—
L_est	0,39	0,20	—	—	-0,32
	(0,00)	(0,09)	—	—	(0,01)

(.) Prob > |R|

Na tabela anterior observa-se que, para os alunos do curso A:

. O desempenho nas disciplinas física, geografia e matemática não mostrou relação significativa com as variáveis categóricas observadas.

. Há evidência de associação entre o desempenho em português, biologia e química e as variáveis X_1, X_3 e X_4. Isto aponta melhores desempenhos, nestas disciplinas, dos alunos que estudaram primeiro grau em escolas privadas, e segundo grau em curso comum e diurno.

. O desempenho em português também apresentou associação X_2, o tipo de estabelecimento de ensino no segundo grau: melhores desempenhos associados a segundo grau predominantemente em escolas particulares.

. A disciplina língua estrangeira (L_est), apresentou associação com X_1, X_2 e X_5 ou seja com o tipo de estabelecimento de ensino no primero e segundo grau e com ocupação do pai. Epera-se melhor desempenho dos alunos de escolas particulares e com a ocupação do pai nos estratos superiores.

. O desempenho na história mostrou associação significativa com a variável X_4, período de estudo de segundo grau. A evidência aponta melhores desempenhos dos alunos que estudaram em período diurno.

ii) Curso B

. Correlação entre variáveis contínuas

Tabela A.11 - Correlação entre as variáveis contínuas no curso B, escolhidas para exame.

	Matérias					
Matéria	Port	Biol	Quim	Hist	Fis	Geog
Quim	—	0,38	—	—	—	—
	—	(0,00)	—	—	—	—
Hist	0,23	—	—	—	—	—
	(0,03)	—	—	—	—	—
Fis	—	0,26	0,35	—	—	—
	—	(0,02)	(0,00)	—	—	—
Geog	—	0,22	—	0,28	-0,23	—
	—	(0,04)	—	(0,01)	(0,03)	—
Mat	—	—	0,28	—	0,58	-0,31
	—	—	(0,01)	—	(0,00)	(0,00)

(.) Prob > |R|

A tabela anterior revela que no curso B, o desempenho nas

disciplinas de exatas apresentou associação positiva entre física, química e biologia (tomadas duas a duas) entre as disciplinas matemática, física e química (quando consideradas duas a duas) porém não há evidência de relação linear entre matemática e biologia.

Na área de humanas encontrou-se evidência de relação linear entre as notas em história e geografia e também entre história e português. Note-se que o desempenho em língua estrangeira não apresentou associação com outras disciplinas da área.

Geografia é a única matéria de humanas que mostrou associação linear significativa com as disciplinas de exatas: positiva com biologia e negativa com matemática e física.

. Correlação entre variáveis categóricas

Tabela A.12 - Correlação entre as variáveis categóricas no curos B escolhidas para exame.

	Variáveis Categóricas			
Variáveis Categóricas	X_1	X_2	X_3	X_4
X_2	0,56 (0,00)			
X_3	0,24 (0,02)	0,52 (0,00)		
X_5	-0,21 (0,04)	— —	— —	— —
X_6	— —	-0,21 (0,05)	-0,28 (0,01)	0,19 (0,08)

(.) Prob > |R|

. A tabela A.12 mostra associação entre as variáveis X_1 e X_2 que representam tipo de estabelecimento de ensino no primeiro e no segundo grau, respectivamente. A evidência aponta uma tendência a manter o tipo de escola, seja pública ou particular, durante todo o período escolar.

. O tipo e escola (variáveis X_1 e X_2), também se relacionou com X_3, curso de segundo grau assim, a tendência é que os alunos de escolas públicas fazem cursos técnicos enquanto que alunos de escolas particulares fazem cursos comuns.

. A variável X_6, participação na vida econômica da família, apresentou associações significativas com o tipo de escola, período e curso de segundo grau (X_2, X_4 e X_2, respectivamente). As evidências sugerem que alunos de escolas públicas de segundo grau tendem a fazer cursos técnicos em período noturno e também tendem a trabalhar.

. Por último, a ocupação do pai X_5, se relaciona com X_1, o tipo de escola que predominou no primeiro grau: escolas particulares se associam com ocupações dos estratos superiores e escolas públicas com ocupações dos estratos inferiores.

. **Correlação entre variáveis contínuas e categóricas**

Tabela A.13 - Correlação entre variáveis contínuas e categóricas no curso B, escolhidas para exame.

Matérias	Variáveis categóricas		
	X_1	X_5	X_6
Português	0,22	—	—
	(0,04)	—	—
Biologia	—	—	-0,19
	—	—	(0,07)
Química	—	-0,21	-0,26
	—	(0,04)	(0,01)
História	—	-0,18	—
	—	(0,10)	—
Língua estrangeira	—	-0,19	—
	—	(0,07)	—

(.) Prob > |R|

Foi encontrada uma associação significativa entre o desempenho em português e X_1, tipo de estabelecimento de ensino no primeiro grau (melhor desempenho associado a escolas particulares).

As notas em biologia apresentaram associação com X_6, participação na vida econômica da família: notas mais altas associadas a alunos que não trabalham.

A variável X_5, ocupação do pai, mostrou associação com o desempenho em química, história e língua estrangeira. A evidência aponta melhores desempenhos associados a ocupação do pai nos estratos superiores.

iii) Curso C

. Correlação entre as variáveis contínuas

Tabela A.14 - Correlação entre as notas no curso C escolhidas para exame.

	Matérias				
	Port	Biol	Quim	Fis	Mat
Quim	-0,33 (0,07)	0,57 (0,00)			
Hist	— —	0,37 (0,05)	— —		
Fis	— —	0,43 (0,02)	0,41 (0,02)		
Mat	— —	— —	— —	0,33 (0,07)	
L_est	— —	— —	— —	-0,32 (0,08)	-0,34 (0,06)

(.) Prob > |R|

Na tabela anterior observa-se que na área de exatas há evidência de relação linear positiva entre o desempenho em biologia, química e física (consideradas duas a duas). Também há relação significativa entre as notas de física e matemática porém, o desempenho em matemática não apresenta associação significativa com o desempenho em química ou biologia.

As disciplinas de humanas não apresentaram correlações significativas entre elas.

Por último, considerando áreas diferentes, foram observadas as

seguintes associações significativas entre as disciplinas:

- relação negativa entre português e química,
- relação positiva entre história e biologia e
- relação negativa entre língua estrangeira e matemática e língua estrangeira e física.

. Correlação entre as variáveis categóricas

Tabela A.15 - Correlação entre as variáveis categóricas no curso C escolhidas para exame.

Variáveis Categóricas	Variáveis Categóricas		
	X_1	X_2	X_3
X_2	0,46 (0,01)		
X_3	0,31 (0,09)	0,41 (0,02)	
X_4	— —	-0,36 (0,05)	— —
X_6	— —	— —	-0,58 (0,00)

(.) Prob > |R|

A tabela A.15 mostra que as respostas dos alunos do curso C às variáveis categóricas apresentaram as seguintes relações significativas:

. Associação entre X_1 e X_2 que indica o tipo de estabelecimento de ensino tende a ser mantido durante todo o período escolar.

. Associação entre tipo de estabelecimento de ensino no primeiro e segundo grau (X_1, X_2) e curso de segundo grau (X_3). A tendência é

que os alunos de escolas públicas estudem cursos técnicos e que os alunos de escolas particulares façam cursos comuns.

. O tipo de escola de segundo grau, X_2, apresentou associação com o período de estudo no segundo grau, X_4. A tendência é que os alunos de escolas particulares estudem em período diurno e que os alunos de escolas públicas em noturno.

. Finalmente, encontrou-se associação significativa entre X_3 e X_6, ou seja, entre o curso de segundo grau e a participação na vida econômica da família. As evidências indicam que os alunos de escolas técnicas tendem a trabalhar.

. Correlação entre as variáveis contínuas e categóricas

Tabela A.16 - Correlações entre as variáveis contínuas e categóricas no curso C escolhidas para exame.

	Variáveis Categóricas				
Matérias	X_1	X_2	X_3	X_5	X_6
Biologia	—	—	0,31	—	-0,39
	—	—	(0,09)	—	(0,03)
Química	—	—	0,46	—	-0,37
	—	—	(0,01)	—	(0,04)
História	—	0,33	0,39	—	—
	—	(0,07)	(0,03)	—	—
Matemática	—	—	—	0,35	—
	—	—	—	(0,05)	—
Língua estrangeira	0,34	—	—	—	—
	(0,06)	—	—	—	—

(.) Prob > |R|

O desempenho em biologia e química apresentou associação com as variáveis X_3 e X_6 que representam curso de segundo grau e participação na vida econômica da família respectivamente. Espera-se melhores desempenhos, nestas matérias, dos alunos de escolas comuns e que não trabalham.

As notas em história também apresentaram relação significativa com X_3, curso de segundo grau: melhores desempenhos em história associados a alunos que terminaram em curso comum.

O desempenho em matemática revelou associação significativa com a ocupação do pai. As evidências apontam bom desempenho associado a ocupação nos estratos inferiores.

REFERÊNCIAS BIBLIOGRÁFICAS

Afifi, A. A. e Elashoff R. M. (1969). Multivariate two sample tests with dichotomous and continuous variables. 1. The location model. *Ann. Math. Stat.*, 40, 1, 290-298.

Aitchinson, J. e Aitken, C. G. G. (1976). Multivariate binary discrimination by the kernel method. *Biometrika*, 63, 3, 413-420.

Albert, A. (1978). *Quelques apports nouveaux à l'analyse discriminante.* Tese de Ph. D., Faculté des Sciences, Université de Liège, Liège, France.

Anderson, J. A. (1972). Separate sample logistic discrimination. *Biometrika*, 59, 19-35.

Anderson, J. A. (1974). Diagnosis by logistic discrimination function: further practical problems and results. *Appl. Stat.*, 23, 397-404.

Anderson, J. A. (1975). Quadratic logistic discrimination. *Biometrika*, 62, 149-154.

Anderson, J. A. (1979). Multivariate logistic compounds. *Biometrika*, 66, 17-26.

Anderson, J. A. (1982). Logistic Discrimination. *Em* P.R. Krishnaiah e L. Kanal (Eds.), *Handbook of Statistcs.* Vol. II. *Classification, Pattern Recognition an Reduction of Dimension*, pp. 169-191, North-Holland: Amsterdam.

Anderson, J. A. e Blair, V. (1982). Penalized maximum likelihood estimation in logistic regression and discrimination. *Biometrika*, 69, 123-136.

Anderson, J. A. e Richardson, S. C. (1979). Logistic discrimination and bias correction in maximum likelihood estimation. *Technometrics*, 21, 71-78.

Anderson, T. W. (1958). *Introduction to Multivariate Statistical Analysis*. Wiley: New York.

Breiman, L. Meisel; W. e Purcell, E. (1977). Variable estimates of multivariate densities. *Technometrics*, 19, 135-144.

Cacoullos, T. (1966). Estimation of multivariate densitiy kernel method *Ann. Inst. Stat. Math.*, 18, 179-189.

Cacoullos, T. (Ed.) (1973). *Discriminant Analysis and Applications*. Academic Press: New York.

Chang, P. C. e Afifi, A. A. (1974). Classification based on dichotomous and continuous variables. *J. Am. Stat. Assoc.*, 69, 336-339.

Cochran, W. G. e Hopkins, C. (1961). Some classification methods with multivariate qualitative data. *Biometrics*, 17, 10-32.

Cornfield, J. (1962). Joint dependence of risk of coronary heart disease on serum cholesterol and systolic blood presure: A discriminant function analysis. *Fed. Proc.*, 21, 58-61.

Cover, T. M. e Hart, P. E. (1967). Nearest neighbour pattern classification. *IEEE Trans. Inf. Theory*, IT-13, 21-27.

Cox, D. R.(1966). Some procedures associated with the logistic qualitative response curve. *Em* F. N. David (Eds.), *Research papers in Statistics: Festschrift for J. Neyman*, pp. 55-71. Wiley New York.

Cox, D. R. e Hinkley, D. V. (1974). *Theoretical Statistics*. Chapman and Hall: London.

Daudin, J. J. (1986). Selection of variables in mixed-variable discriminant analysis. *Biometrics*, **42**, 473-481.

Day, N. E. e Kerriedge, D. F. (1967). A general maximum likelihood discriminate. *Biometrics*, **23**, 313-323.

Deming, M. E. Stephan, F. (1940). On a least square adjustment of a sampled frequency table when expected marginal totals are known. *Ann. Math. Stat.*, **11**, 427-444.

Devroye, L. e Wagner T. J. (1982). Nearest Neighbor Methods in Discrimination. *Em* P.R. Krishnaiah e L. Kanal (Eds.), *Handbook of Statistcs*. Vol. II. Classification, Pattern Recognition and Reduction of Dimension, pp. 193-197, North-Holland: Amsterdam.

Dolby, J. L. (1970). Some statistical aspects of character recognition. *Technometrics*, **12**, 231-245.

Duda, R. O. e Hart P. E. (1973). *Pattern Classification and Scene Analysis*. Wiley: New York.

Fienberg, S. E. (1970). An iterative procedure for estimation in contingency tables. *Ann. Math. Stat.*, **41**, 907-917.

Fisher, R. A. (1936). The use of multiple measurement in taxonomic problems. *Ann. Eug.*, **7**, 179-188.

Gill, P. E. e Murray, W. (1972). Quasi-Newton methods for unconstrained optimisation. *J. Inst. Math. Appl.*, 9, 91-108.

Goldstein, M. e Dillon, W. R. (1978). *Discrete Discriminant Analysis*. Wiley: New York.

Gower, J. C. (1966). Some distance properties of latent root and vector methods used in multivariate analysis. *Biometrika*, **53**, 325-338.

Habbema, J. D. F., Hermans, J. e Remme, J. (1978). Variable kernel density estimation in discriminat analysis. *Em* L. C. A. Corsten e J. Hermans (Eds.), *Compstat 1978*, pp 178-185. Physica-Verlag: Vienna.

Habbema, J. D. F., Hermans, J. e van den Broek, K. (1974). A Stepwise discriminant analysis program using density estimation. *Em* G. Brukmann, F. Ferschl e L. Schmetterer (Eds.), *Compstat 1974*, pp. 101-110. Physica-Verlag: Vienna.

Haberman, S. T. (1972). Log-linear fit for contingency tables. Algorithm AS51, *Appl. Stat.*, **21**, 2, 218-225.

Hall, P. (1981). On nonparametric multivariate binary discrimination. *Biometrika*, **68**, 287-294.

Hermans, J., Habbema, J. D. F., Kasanmoentalib T. K. D. e Raatgever, J. W. (1984). ALLOC80 discriminant analysis program.

Huberty, C. J. (1975). Discriminant analysis. *Rev. Educ. Res.*, **45**, 543-593.

James, B. R. (1981). *Probabilidade: Um curso a nível intermediário.* Livros Técnicos e Científicos: Rio de Janeiro.

Jhon, S. (1960). On some classification problems. *Sankhya*, **22**, 301-308.

Jhon, S. (1963). On classification by statistics R and Z. *Ann. Inst. Stat. Math.*, **14**, 237-246.

Johnson, N. L. e Kotz, S. (1969). *Discrete Distributions.* Wiley: New York.

Jones, R. H. (1975). Probability estimation using a multinomial logistic function. *J. Stat. Compt. Simul.*, **3**, 315-329.

Knoke, J. D. (1982). Discriminat analysis with discrete and continuous variables. *Biometrics*, **38**, 191-200.

Krusinska, E. (1989). New procedure for selection of variables in location model for mixed variable discrimination. *Biom. J.*, **31**, 551-523.

Krzanowski, W. J. (1971). The algebric basis of classical multivariate methods. *Statistician*, **20**, 51-61.

Krzanowski, W. J. (1975). Discrimination and classification using both binary and continuous variables. *J. Am. Stat. Assoc.*, 70, 782-790.

Krzanowski, W. J. (1976). Canonical representation of the location model for discrimination or classification. *J. Am. Stat. Assoc.*, 71, 845-848.

Krzanowski, W. J. (1977). The performance of Fishers's linear discriminant function under non-optimal conditions. *Technometrics*, 19, 191-200.

Krzanowski, W. J. (1979). Some linear tramsformations for mixtures of binary and continuous variables, with particular reference to discriminant analysis. *Biometrika*, 66, 33-39.

Krzanowski, W. J. (1980). Mixtures of continuous and categorical variables in discriminant analysis. *Biometrics*, 36, 493-499.

Krzanowski, W. J. (1982). Mixtures of continuous and categorical variables in discriminant analysis: A hypothesis-testing approach. *Biometrics*, 38, 991-1002.

Krzanowski, W. J. (1983a). Stepwise location model choice in mixed-variable discrimination. *Appl. Stat.*, 32, 260-266.

Krzanowski, W. J. (1983b). Distance between populations using mixed continuous and categorical variables. *Biometrika*, 70, 235-243.

Krzanowski, W. J. (1986). Multiple discriminant analysis in the presence of mixed continuous and categorical data. *Comp. & Math. Appls.*, 2, 179-185.

Lachenbruch, P. A. (1967). An almost unbiased method of obtaining confidence intervals for the probability of missclassification in discriminant analysis. *Biometrics*, 23, 639-645.

Lachenbruch, P. A. (1975). *Discriminant Analysis*. Hafner: New York.

Lachenbruch, P. A. e Goldstein, M. (1979). Discriminant analysis. *Biometrics*, 35, 69-85.

Leung, C. Y. (1989). The studentized location linear discriminant function. *Comm. Stat.-Theory Meth.*, **18**, 3977-3990.

Mardia, K. V. Kent. J. T. e Bibby, J. M. (1979). *Multivariate Analysis*. Academic Press: London.

Murray, G. D. e Titterington. (1978). Estimation problems with data from a mixture. *Appl. Stat.*, **27**, 325-334.

Murthy, V. K. (1966). Nonparametric estimation of multivariate densities with applications. *Em.* P.R. Krishnaiah (Ed.). *Multivariate Analysis*, pp. 43-56. Academis Press: New York.

Olkin, F. e Tate, R. F. (1961). Multivariate correlation models with mixed discret and continuous variables. *Ann. Math. Stat.*, **32**, 448-465.

Press, S. J. e Wilson, S. (1978). Choosing between logistic regression and discriminant analysis. *J. Am. Stat. Assoc.*, **73**, 699-705.

Remme, J., Habbema, J. D. F. e Hermans, J. (1980). a simulative comparison of linear, quadratic and kernel discrimination. *J. Stat. Compt. Simul.*, **11**, 87-106.

SAS Instituite, Inc. (1989). *SAS/STAT User's Guide*: Versão 6, quarta edição Vol. 1 e 2. SAS Intitute, Inc.: Cary, N. C.

Seber, G.A. F. (1984). *Multivariate Observations*. Wiley: New York.

Tate, R. F. (1954). Correlation between a discrete and continuous variables. *Ann. Math. Stat.*, **25**, 603-607.

Toussaint, G. T. (1974). Bibliography on estimation of misclassification. *IEEE Trans. Inf. Theory* **IT-20**, 472-479.

Tu, C.-T. e Han, C.-P. (1982). Discriminant analysis based on binary and continuous variables. *J. Am. Stat. Assoc.*, **77**, 447-454.

Vlachonikolis, I. G. (1985). On the asymptotic distribution of the location linear discriminant function. *J. R. Stat. Soc. B.*, **47**, 498-509.

Vlachonikolis, I. G. (1986). On the estimation of the expected probability of misclassification in discriminant analysis with mixed binary and continuous variables. *Comp. & Math. Appls.*, 2, 187-195.

Vlachonikolis, I. G. (1990). Predictive discrimination and classification with mixed binary and continuous variables. *Biometrika*, 77, 657-662.

Vlachonikolis, I. G. e Marriott, F. H. C. (1982).Discrimination with mixed binary and continuous data. *Appl. Stat.*, 31, 23-31.

Wojciechowski, T. J. (1987). Nearest neighbor classification rule for mixtures of discrete and continuous random variables. *Biom. J.*, 8, 953-959.