SELEÇÃO DE VARIÁVEIS BINÁRIAS PARA DIAGNÓSTICO MÉDICO: COMPARAÇÃO DO CRITÉRIO DE KOKOLAKIS COM OUTROS CONCORRENTES

Este exemplar corresponde a redação final da tese devidamente corrigida e defendida pelo Sr. Adão Luiz Hentges e aprovada pela Comissão Julgadora.

Campinas, 19 de julho de 1989

Prof. Dr. José Antônio Cordeiro
Orientador

Dissertação apresentada ao Instituto de Matemática, Estatistica e Ciência da Computação, UNICAMP, como requisito parcial para obtenção do Titulo de Mestre em Estatística.

"Aos trabalhadores brasileiros, que financiaram este estudo"

Agradecimentos:

A José Antônio Cordeiro pelo apoio recebido e pela amizade que norteou nosso contato;

A Belmer Garcia Negrillo e Carlos Alberto de Bragança Pereira pelo auxílio e pela disponibilidade, que poderia ter sido melhor aproveitada por minha parte;

Aos professores e funcionários do IMECC;

A Rose e Nete do xerox do CAMECC;

A Iara Rehder pela ajuda valiosa que possibilitou a consecução da edição deste trabalho;

Aos amigos do curso e do futebol pelo convívio e momentos de descontração;

Às psicólogas Marilda Lipp e Maria José Nery pela possibilidade de trabalhar com dados reais sobre o problema do "stress";

Às forças do Universo, que colocaram a Rajo em meu caminho.

INTRODUÇÃO

Este trabalho se constitui numa tentativa inicial de compreender o processo de construção das regras de diagnóstico médico, especialmente no que diz respeito à seleção das variáveis para tal propósito.

Obviamente esta tarefa é muito complexa, exigindo uma relação harmoniosa entre as regras estatísticas de classificação e sua aplicabilidade e clareza por parte dos médicos que as utilizariam. Enquanto o estatístico tenta criar um critério de diagnóstico a partir de modelos e suposições específicas, a maioria dos médicos tem um procedimento já estabelecido para a tomada de decisão sobre uma intervenção no paciente, baseado em seu quadro clínico observado.

Entretanto, esta habilidade que o médico adquiriu ao conviver com inúmeros pacientes expostos à enfermidade de interesse não consegue ser captada inteiramente pelo estatístico. Além de conviver mais intimamente com o problema, o médico utiliza sua percepção e prende-se a detalhes muito ricos, não perfeitamente mensuráveis por um matemático ou possíveis de serem expressos por um indicador médico convencional.

Assim, é natural que as regras estatísticas apresentem grande dificuldade ou mesmo não cheguem a alcançar a utilidade que gostariam. Diversos autores tem discutido este problema em seus trabalhos na área clínica, citando a falha das regras estatísticas em estabelecer critérios pouco claros ou mesmo coerentes a respeito do diagnóstico sobre o paciente. Uma transformação de dados, por exemplo, ou a inclusão de variáveis com qualidades questionáveis de informação, pode conduzir a um modelo confuso e não inteligível por parte dos médicos. Neste caso, embora a regra de classificação tenha propriedades matemáticas satistatórias, provavelmente seria ignorada pelos clínicos.

Outros autores, buscando captar a experiência e informação relevante, tentaram estabelecer modelos de inteligência artificial junto aos hospitais e centros clínicos para tornar a aprendizagem mais dinâmica. Entretanto, as dificuldades para implementar os sistemas computacionais e estabelecer regras mais ágeis também são apontadas.

Mesmo com tais dificuldades, a tentativa de construção de regras que possam eventualmente prevenir um enfarte futuro em um paciente é por demais justificável.

Consideremos então o problema de alocar o paciente a uma classe de doenças $D=\{D_1, D_2\}$, onde D_1 e D_2 seriam exclusivas e exaustivas, a partir de indicadores $X=(X_1,\ldots,X_d)$ observados. A alocação seria o paciente "tem/não tem a doença, irá/não irá se recuperar do estado de coma pós-cirurgia" e assim por diante.

Quanto aos indicadores, disponíveis e/ou tradicionalmente observados em pacientes que buscam consulta médica para certa enfermidade, refletiriam o resultado de um teste clínico, um sintoma ou presença de uma característica e seriam considerados como os mais relevantes para emitir o diagnóstico.

A decisão consiste em escolher os melhores indicadores a serem observados para construir a regra e, baseado nos valores assumidos em um certo paciente, que tipo de diagnóstico deverá ser emitido. Nesta alocação, que poderá alterar o tratamento ou mesmo não prevenir um problema iminente, precisamos considerar as conseqüências e não apenas a probabilidade do diagnóstico ser correto.

Considerando iguais conseqüências aos erros, a alocação do paciente à classe que fornece máxima utilidade esperada levaria a classificar o paciente na classe julgada mais provável. Entretanto, ao assessar a verossimilhança de diferentes classes para um paciente com um quadro clínico complexo, requer-se a investigação da incidência conjunta dos indicadores e doença.

Para contornar o problema, costuma-se assumir independência entre os indicadores, para cada classe fixada da doença. Este procedimento, questionável, pode levar a erros sérios de alocação, de forma que são consideradas estruturas de dependência e mesmo outro tipo de função utilidade para a alocação dos pacientes , o que estabelece novos critérios de seleção dos indicadores.

Também sob outro contexto, são selecionados indicadores de forma sequencial, ou seja, condicionado na observação de $X_i=x_i$ busca-se a observação da variável X_j $(j\neq i)$ mais útil para classificação. Assim,

a partir de um indicador já observado, seriam buscados somente aqueles contendo informação significativa, escapando-se do problema quase intratável de escolher o melhor sub-grupo de variáveis em um grupo disponível muito grande.

Pode-se ainda selecionar os indicadores através dos fatores de Bayes, buscando aqueles com maior diagnosticabilidade, isto é, o poder de modificar os valores das quantidades diagnóstico (probabilidades de ter ou não a doença) a partir da evidência observada.

Aliás, já foi constatada a vantagem de adotar o diagnóstico sob o contexto Bayesiano, mesmo com a natural dificuldade de elaborar a distribuição a priori para os modelos de classificação.

Utilizando conceitos de Teoria da Informação, como a entropia de uma distribuição ou a divergência entre duas populações, pode-se selecionar as variáveis com maior valor para a alocação dos pacientes em um dos grupos.Dentro desta linha será selecionado um sub-grupo fixado de variáveis entre um conjunto disponível através de um coeficiente que meça a utilidade dos indicadores em estabelecer a separação dos grupos.

Este coeficiente, sob enfoque Bayesiano, será comparado com alguns outros quanto ao poder de detectar as melhores variáveis, em estruturas simuladas que permitem a identificação dos testes mais informativos.

A partir das variáveis selecionadas, o desempenho da regra de classificação será medido através da taxa de erro obtida sob várias estruturas, com diferentes configurações de dependência entre os indicadores clínicos.

Em função do interesse inicial sobre diagnóstico médico, que ao final ficou restrito à utilização de variáveis binárias, muitos artigos foram lidos. Embora não haja uma referência explícita a alguns deles no texto, os mesmos se inserem no contexto deste estudo e servem como fonte de consulta a interessados.

1. A CONSTRUÇÃO E ANÁLISE DE DISTRIBUIÇÕES DIAGNÓSTICO

1.1 Objetivo

Seja $D=\{D_1,\ldots,D_t\}$ um conjunto de classes de doenças, onde "doença" é usado como um termo associado a uma enfermidade ou mesmo uma classe de resposta de um paciente frente a determinado tratamento, assumindo-se inicialmente que um paciente pertence a uma e somente uma destas classes. As t classes podem representar também distintas doenças presentes em uma certa clínica especializada ou alternativamente podem ser grupadas juntas em uma classe correspondendo à doenças que exijam tratamento similar.

Em cada paciente um conjunto de indicadores $X=(X_1,\ldots,X_d)'$ é observado, onde "indicador" representa um sintoma observado, o resultado de um teste clínico ou a presença de uma característica no paciente, segundo Good e Card(1971).

Um sistema de diagnóstico estatístico tenta utilizar algum modelo matemático para assessar a quantidade $P(D_i/X)$, $i=1,\ldots,t$, que sumarize o suporte dado a cada classe da doença pela evidência disponível, na qual, se apropriado, uma regra de classificação pode ser baseada.

No momento da alocação do paciente por algum critério, a quantidade $P(D_i/X)$ pode ser interpretada como a estimativa da probabilidade de que este novo paciente esteja em D_i ou então apenas uma medida de evidência, frente aos dados observados.

Spiegelhalter e Knill-Jones(1984), em um artigo muito rico sobre o assunto, apontam como opinião geral na área clínica que a técnica estatística é geralmente muito simplista para problemas reais, inaplicável porque os dados são insuficientes e incompreensível ao usuário. Em resposta parcial apresentam a aplicação de um processo probabilístico a um problema clínico complexo, como o diagnóstico das causas de "dyspepsia". Neste sistema, que procurou a utilização de

técnicas probabilísticas confiáveis, incorporaram-se também alguns aspectos de inteligência artificial à complexidade e explanação .

No histórico que estes autores fazem sobre as técnicas de classificação, evoluindo do uso corrente do teorema de Bayes à técnicas de Inteligência Artificial implementadas nos hospitais, percebe-se a complexidade do problema.

Na discussão que se segue com outros cientistas, conclui-se pela dificuldade em uniformizar critérios de classificação, desde a construção de técnicas que levantam as probabilidades subjetivas junto aos médicos até a obtenção de um desempenho satisfatório dos sistemas computacionais que buscam a sumarização da informação clínica e sua aprendizagem sequencial.

1.2 Paradigmas e Modelos

O *paradigma amostral* concentra-se na distribuição de **X** dado **D**, considerando a variável doença como um parâmetro desconhecido e as facetas(ou indicadores) como variáveis aleatórias com uma distribuição conjunta fixada dado a doença, enfatizando a estimação destas distribuições.

Segundo Dawid(1976), este paradigma surge da situação de sintoma puro, na qual a doença é considerada como causativa. Entretanto, as várias técnicas de diagnóstico derivadas são geralmente aplicadas quando alguns indicadores são apenas classificatórios ou causativos. Além disso, há uma confusão geral da constância de P(**X**/D) sob trocas nas taxas de incidência com constância sob trocas em processos de seleção amostral. Vícios de seleção que surgem naturalmente porque diferentes centros clínicos se especializam em áreas distintas da Medicina e assim podem ter diferentes critérios de admissão dos pacientes e a variação da prevalência da doença em diferentes regiões ou períodos de tempo podem alterar a distribuição de **X** dado D.

Embora o paradigma amostral leve à aplicação de técnicas discriminantes e mais comumente à classificação de Bayes, onde é quase geral assumir que os vários indicadores são condicionalmente independentes dado D, a estimação de P(X/D) geralmente envolve muitos parâmetros, forçando as técnicas ao uso de um número considerável de indicadores.

Alternativamente, o *paradigma diagnóstico* considera a distribuição de **D** dado **X** como o objeto apropriado de análise. Dawid sugere em seu artigo que as vantagens de se trabalhar sob este paradigma são conceituais e práticas.

Percebe-se que sob condições razoáveis os efeitos de seleção podem ser ignorados e a transformação de uma distribuição diagnóstico se movendo de uma população para outra pode ser feita facilmente.

Uma extensão da forma logística de Cox(1968) e Day & Kerridge(1967) para as probabilidades a posteriori com t=2, isto é, $D=\{D_1, D_2\}$,

$$P(D_1/\mathbf{x}) = \exp\{\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d\}\ P(D_2/\mathbf{x})$$

onde,

$$P(D_2/\mathbf{x}) = 1 \,/\, \{1 + \exp(\beta_0 + \beta_1 x_1 + \ldots + \beta_d x_d)\}$$

leva ao modelo logístico

$$P(D_i/\mathbf{x}) = \frac{\exp\{y_i(\mathbf{x})\}}{\sum_j \exp\{y_j(\mathbf{x})\}}, \qquad \text{com} \qquad y_i(\mathbf{x}) = \sum_{q=0}^{d} \beta_{qi} z_q$$

sendo os β's parâmetros desconhecidos e os z's funções especificadas de **x**.

Dawid cita que este modelo surge sempre que a distribuição de **X** dado **D** forma uma família exponencial e os z's atuam como estatísticas suficientes naturais. A grande variedade de situações em que este modelo pode surgir, sua resistência aos efeitos de seleção e o

comportamento simples sob trocas de taxas de incidência torna-o um sério candidato para modelar distribuições diagnóstico.

1.3 Indicadores clínicos

Neste trabalho restringiremos nossa atenção a indicadores binários, tornando-se necessário uma categorização adequada de tal forma a manter relevante a informação de variáveis mensuradas originalmente sob outra escala.

Poderíamos ter, por exemplo, indicadores representando a presença de um sintoma, o resultado de um teste clínico quanto à presença de um vírus, se o paciente fuma mais que 20 cigarros diários ou se sua pressão arterial está abaixo de um certo nível.

No caso de dicotomização de uma variável contínua ou discreta, como o exemplo do consumo de cigarros, seria razoável que um especialista da enfermidade sob estudo é quem defina o ponto de corte (nível crítico de cigarros fumados) para a caracterização.

Entretanto, ao realizar uma análise exploratória num conjunto de dados contendo as características e o verdadeiro estado do paciente quanto à doença, o estatístico pode estabelecer sua própria categorização. Por algum critério, trabalhando com diferentes categorizações das variáveis, poder-se-ia chegar a uma categorização mais informativa, mantendo-se em mente que a mesma possa ser observada com facilidade em novos pacientes e centros clínicos.

Genericamente teríamos

$$X_i = \begin{cases} 1, \text{ se a } i\text{-ésima característica está presente no paciente} \\ 0, \text{caso contrário} \end{cases}$$

Consideraremos, segundo Dawid(1976), que nem **D** nem **X** variem significativamente para um dado indivíduo e que os mesmos podem ser determinados sem erro.

Teremos então um particular par (D_i, x) como realizações das variáveis (D,X), com uma distribuição conjunta sobre a população.

1.3.1 Do relacionamento entre os indicadores

A busca da verossimilhança das diferentes classes, para um paciente com um particular quadro de indicadores, requer a investigação da incidência conjunta da doença e indicadores. Em geral, como estas distribuições e incidências são desconhecidas, é necessária a existência de um conjunto de dados de pacientes antigos com registros disponíveis.

Para contornar a dificuldade em obter as estimativas das possíveis distribuições conjuntas apropriadas em um número muito grande de indicadores disponíveis, assume-se independência condicional entre os indicadores dentro de cada classe da doença.

Segundo Teather(1974), esta suposição leva a considerar apenas a incidência marginal de cada indicador em cada classe, acessando-se apenas a utilidade dos indicadores através de tabelas de duas entradas para (D, X_i).

A tabela abaixo mostra um exemplo hipotético que conduz a erro, baseado em 100 pacientes observados sob dois indicadores dicotômicos.

	X_1		X_2		X_1X_2			
	1	0	1	0	11	10	01	00
$D=D_1$	30	30	30	30	30	0	0	30
$D=D_2^1$	20	20	20	20	0	20	20	0

Cada variável sozinha não fornece informação para discriminação. Entretanto, a distribuição conjunta para (D, X_1, X_2) mostra que as duas variáveis juntas permitem perfeita discriminação entre as duas categorias (tem/não tem a doença).

Além disso, segundo Teather, ao assumir independência não se permite a exploração da redundância da informação gerada pelas variáveis, ou seja, duas ou mais variáveis podem dar a mesma informação em forma diferente sem que isto seja utilizado.

Será discutido adiante, por outro lado, o efeito da dependência entre os indicadores quanto ao critério utilizado para expressar o desempenho das regras de classificação.

1.3.2 Da seleção dos indicadores

A maior questão, segundo Spiegelhalter e Knill-Jones(1984), parece ser se uma entidade complexa sobre a qual se tenha conhecimento incompleto, se o processo de julgamento de um clínico ou o próprio processo da doença, é melhor modelado por um sistema que é incrivelmente grande mas essencialmente determinístico, ou que é parcimonioso e probabilístico.

Como a teoria estatística e prática enfatizam a necessidade para parcimônia nos modelos, os autores acreditam que se a intenção do sistema é boa predição de futuros casos, o enfoque probabilístico é apropriado.

Teather alertou que se desejarmos incluir todas as variáveis disponíveis em uma regra de classificação, o conjunto de dados tomado como base não providenciaria estimativas razoáveis das incidências conjuntas.

Além disso, o custo de aplicação de testes clínicos e a necessidade de rapidez na obtenção dos indicadores para emitir o diagnóstico sobre o paciente são motivos naturais que levam a que se preste atenção somente nas variáveis relevantes e informativas, daí ser necessário o estabelecimento de critérios de seleção das mesmas.

Os critérios que estabelecem a seleção de indicadores para construir a regra dependem basicamente da utilidade L_{ji} de alocar um

paciente à classe D_j quando sua verdadeira classe é D_i.

Se estamos frente a um problema onde seja razoável tratar os diferentes erros possíveis de diagnóstico igualmente, podemos adotar uma estrutura de utilidade mais simples, com utilidade 1 para correta alocação e 0 em caso de erro.

Assim, dado o acesso à probabilidade das diferentes classes, a alocação do paciente à classe que fornece máxima utilidade esperada leva à maximização da chance de diagnóstico correto, alocando-se o paciente à classe julgada mais possível.

Entretanto, tal matriz utilidade será inapropriada em muitos problemas reais.

Bernardo e Bermudez(1985), embora estudando a intenção de voto eleitoral, consideram a seleção de variáveis através de funções utilidade logarítmica e quadrática, onde a conseqüência da classificação pode ser local, isto é, a função utilidade se relaciona com a distribuição diagnóstico apenas com a probabilidade atribuída à verdadeira classe, ou não. Esta condição de regra local pode se tornar vaga para estudos com apenas duas classes a serem diagnosticadas.

O critério estabelecido pelos autores, sob enfoque Bayesiano, seleciona as variáveis que minimizam a entropia esperada ou maximizam a norma quadrática esperada da distribuição diagnóstico resultante, conseguindo-se boas aproximações tanto para estudos prospectivos como retrospectivos. Assim, o método de busca seleciona ou retira um indicador a partir de alteração significativa na função utilidade.

Utilizando-se conceitos de Teoria da Informação, a busca das melhores variáveis conduz àquelas que tenham o maior poder de estabelecer a separação dos grupos com sua distribuição conjunta. Alguns coeficientes e medidas de informação serão estudados adiante, permitindo-se a seleção de variáveis sob tal enfoque.

Stanish e Allred(1981) selecionam variáveis categóricas utilizando a função de entropia como uma medida de variabilidade. O critério usa o teste da razão de verossimilhança(TRV) que, para a

hipótese sob consideração, é idêntico às estatísticas de informação de mínima discriminação.

A seleção é baseada em sua contribuição à redução de entropia na variável dependente, sendo a significância assessada pelo TRV mas com valor crítico associado ao nível de significância geral, fornecendo um método que tende a selecionar mais variáveis.

Havendo um grande número de variáveis disponíveis, a seleção de um sub-grupo de tamanho k fixado pode se tornar impraticável devido ao número enorme de sub-grupos a considerar. Além disso, a seleção do melhor sub-grupo de $(k+1)$ variáveis poderia não conter o sub-grupo já selecionado de tamanho k.

Assim, são desenvolvidos métodos sequenciais de seleção de variáveis, incluindo-se uma nova variável ao sub-grupo com o qual já se trabalha a partir de algum critério que mede se sua contribuição é significativa. Para a maioria deles, entretanto, o "ponto de corte" para definir a importância da variável é discutível, em função do problema estudado.

Preocupados com a identificação de categorias de resposta específicas que justificam mensuração, Goldstein e Dillon(1977) utilizam a medida de divergência entre dois grupos segundo o critério de Kullback(1959).

Através da divergência máxima sucessiva $\hat{J}(X_1)$, $\hat{J}(X_1,X_2)$,... e das divergências condicionais $\hat{J}(X_2/X_1=x_1)$, $x_1=0,1$, são selecionadas as variáveis mais significativas baseado na distribuição assintótica de κ^2 (qui-quadrado).

Entretanto, este ajuste do valor crítico por κ^2 pode levar à seleção de mais variáveis do que poderia ser necessário com base nas propriedades estritas da distribuição das estatísticas $max\ \hat{J}$, sob o qual o critério de seleção destes autores trabalha.

Um indicador clínico pode ser escolhido também considerando sua habilidade em alterar as probabilidades a priori de ter ou não ter a doença, isto é, $P(D_1)$ e $P(D_2)$, onde D_2 é o complementar $\bar{D}_1$ de D_1, especificadas inicialmente por um médico em função de sua experiência.

Este tipo de análise leva ao estudo do *peso da evidência*, isto é, o logaritmo do fator de Bayes (definido como a razão do "odds" a posteriori para o "odds" a priori). Considerando uma única variável, o fator de Bayes em favor de D_1, a partir da evidência observada X=1, seria

$$\frac{[P(D_1/X=1)/P(D_2/X=1)]}{[P(D_1)/P(D_2)]}$$

(igual a *sensibilidade* dividida por 1-*especificidade*).

Pereira e Pericchi(1985) sugerem a seleção do indicador apropriado a partir do máximo poder de diagnóstico, definido por uma medida de divergência, estabelecida sobre o *peso da evidência*. Neste trabalho são obtidos os momentos para o estimador do *peso da evidência*, para uma ou mais variáveis.

Pode-se considerar também as funções

$$f_{(i,j)}(\delta) = P(D=D_1/X_1=i, X_2=j) - \delta$$
$$f_{(i,.)}(\delta) = P(D=D_1/X_1=i) - \delta$$
$$f_{(.,j)}(\delta) = P(D=D_1/X_2=j) - \delta$$

onde i,j=0,1 e $\delta=P(D=D_1)$, para buscar o(s) teste(s) que mais contribui para alterar a quantidade δ especificada a priori para D_1, o que é estudado por Pereira e Barlow(1989).

Em outro trabalho sobre o assunto, mas com enfoque ligeiramente diferente, Pereira(1989) obtém as distribuições a posteriori da sensibilidade e da especificidade, as distribuições preditivas e as probabilidades diagnóstico, expressando o desenvolvimento dos resultados através de diagramas.

1.4 Das regras de classificação

1.4.1 O desempenho das regras

A partir do acesso à $P(D_i/X)$ através de alguma técnica apropriada, sua utilização pode consistir numa informação auxiliar ao médico que trata do paciente ou numa tomada de ação quanto ao tratamento. Evidentemente o tipo de doença e o propósito que motivou a busca da quantidade diagnóstico é que definem sua utilização.

Especificamente no caso de alocação, para um paciente classificado sob uso da regra em função de $P(D_i/X)$, teríamos

	Regra	
Paciente	É portador	Não é
Tem doença(D_1)	correto	"não previniu"
Não tem (D_2)[1]	"alarme falso"	correto

É óbvio que os erros de alocação deverão ter tratamento diferenciado em função da gravidade da doença investigada, considerando-se uma tomada de decisão efetiva sob tal alocação, como decidir por cirurgia, alterar ou interromper tratamento, etc. Assim, uma função utilidade adequada deverá atribuir as perdas proporcionais a um "alarme falso" (classificar o paciente como doente quando na verdade ele é sadio) ou "não prevenção" (classificar o paciente como livre de um perigo de saúde a partir das evidências observadas, quando o mesmo está prestes a tê-lo).

Torna-se necessário também, na utilização da função utilidade, a devida caracterização da classe de doenças a ser diagnosticada.

Embora seja insensível à seriedade das conseqüências dos diferentes erros possíveis de alocação, a medida de separação mais comumente utilizada é a proporção de erros cometidos nos dados sob teste alocados a uma classe errada.

Titterington *et. al.*(1981) descrevem várias medidas de separação e de confiabilidade para as regras de diagnóstico, citando também a construção de uma matriz-perda associada a erros de classificação de pacientes que após permanecerem em coma por mais de 6 horas poderiam ter morte ou vida vegetativa, problemas severos ou recuperação.

Teather(1974) estabelece uma medida de acurácia do verdadeiro diagnóstico, ou incidência de correto diagnóstico, encontrando sua distribuição a posteriori. A partir destes resultados pode-se construir intervalos Bayesianos de confiança para a medida criada, tanto para uma variável como para a inclusão de um novo indicador.

Acreditamos que somente o problema real que está sendo considerado é que poderá garantir a validade de um certo critério para avaliar se a regra de classificação é boa.

Embora seja trabalhoso e resultado de muita consulta junto à especialistas que convivem com a enfermidade analisada, parece ser fundamental a construção de uma matriz-perda adequada, o que poderá levar à seleção de variáveis e classificação dos pacientes dentro do contexto de Teoria da Decisão.

1.4.2 A validade da regra construída

Segundo Spiegelhalter e Knill-Jones, um número de dificuldades técnicas surge quando tentamos utilizar uma regra de classificação em uma população diferente daquela onde a regra foi construída.

Em primeiro lugar, diferentes dados clínicos podem ser rotineiramente coletados, de forma que somente variáveis comumente disponíveis deveriam ser usadas para classificação. Além disso, se a variação do observador em sugerir indicadores difere no novo local, isto pode viciar as predições probabilísticas.

A prevalência das doenças também pode variar de um local para outro, devido à variação geográfica natural ou a diferentes razões de encaminhamento dos pacientes ao centro clínico.

Além de processos intrínsecos à própria doença, sua presença pode variar devido a diferentes definições dos indicadores.

Dawid(1976) considera o vício de seleção decorrente de diferentes critérios de admissão de pacientes aos centros clínicos, levando o grupo estudado a não ser representativo da população geral.

Ao supor que nenhuma outra informação é de relevância para diagnóstico quando os indicadores diagnósticos **X** já são conhecidos e considerar a distribuição conjunta de (**X**,D) não variando com tempo ou locação, o autor estabelece que as distribuições diagnóstico não são afetadas seriamente por este tipo de erro.

Além da preocupação com a *transferibilidade* da regra de classificação, é válido preocupar-se também com a aquisição de conhecimento, o que é possível através do convívio com especialistas frente a um grande número de informações e problemas práticos que surgem a sua volta.

Spiegelhalter e Knill-Jones apontam o sucesso e dificuldades encontradas por várias tentativas de implementar diferentes sistemas computacionais que trabalham com Inteligência Artificial em clínicas e hospitais.

2. TAXAS DE ERRO PARA VARIÁVEIS BINÁRIAS SOB A REGRA DE BAYES

Geralmente os métodos desenvolvidos para análise discriminante linear com variáveis normais são simplesmente transferidos ao caso de variáveis qualitativas. Entretanto, este tipo de variáveis tem qualidades especiais que precisam ser consideradas, como por exemplo, sua estrutura de interação não é inteiramente descrita com interações de primeira ordem, como é o caso com variáveis normais, ou seja, somente as correlações entre cada par de variáveis qualitativas não determina seu comportamento conjunto.

Para selecionar sub-grupos de variáveis qualitativas em uma maneira ótima somente variáveis qualitativas dicotômicas serão consideradas, com os parâmetros de probabilidade das variáveis assumidos como conhecidos, por questão de simplicidade.

Consideremos o problema já exposto, ou seja, alocar o paciente à classe D_1 ou D_2, onde $D_2=\bar{D}_1$, a partir de um vetor observado de indicadores $X=(X_1,\ldots,X_d)'$. Se tratarmos as conseqüências dos diferentes erros possíveis de diagnóstico com igual peso, isto é, perda $l(D,a(\mathbf{x}))=0$ quando o paciente é alocado corretamente e perda igual a 1 quando a alocação é errônea(independentemente do fato de ser "falso positivo" ou "falso negativo"), a alocação que fornece o máximo valor esperado para esta utilidade conduz o paciente à classe mais provável. Desta forma, chega-se à aplicação da *regra de classificação de Bayes*.

Adotando a notação

Π_1: prevalência populacional de indivíduos com a doença, isto é, $P(X\in D_1)$

P_1: probabilidade *a priori* atribuída à prevalência da doença na população através da informação médica, isto é, $P(D_1)$

$P_2=1-P_1$

distribuição unidimensional

$P(X_j=1/D_1)=p_j \qquad (j=1,\ldots,d)$

$P(X_j=1/D_2)=q_j$

e para uma distribuição bidimensional de (X_i, X_j)

$$P(X_i=x_i, X_j=x_j/D_1)=p_{x_i x_j} \qquad x_i, x_j=0,1$$
$$P(X_i=x_i, X_j=x_j/D_2)=q_{x_i x_j} \qquad i,j=1,\ldots,d$$

as distribuições de probabilidade de (X_i, X_j) em D_1 e D_2, ao assumir-se independência condicional das d variáveis nos dois grupos, podem ser escritas como

$$p_{11}=p_i p_j \qquad q_{11}=q_i q_j$$
$$p_{10}=p_i(1-p_j) \qquad q_{10}=q_i(1-q_j)$$
$$p_{01}=(1-p_i)p_j \qquad q_{01}=(1-q_i)q_j$$
$$p_{00}=(1-p_i)(1-p_j) \qquad q_{00}=(1-q_i)(1-q_j)$$

Assumindo-se que Π_1 é conhecida, isto é, $P_1=\Pi_1$, a regra ótima de classificação de Bayes é dada por:

$$\begin{array}{ll} & \Pi_1 f(\mathbf{x}/D_1) > \Pi_2 f(\mathbf{x}/D_2): \text{ alocar o paciente na classe } D_1 \\ \text{Se} & \Pi_1 f(\mathbf{x}/D_1) = \Pi_2 f(\mathbf{x}/D_2): \text{ alocar o paciente aleatoriamente em } D_1 \text{ ou } D_2 \\ & \Pi_1 f(\mathbf{x}/D_1) < \Pi_2 f(\mathbf{x}/D_2): \text{ alocar o paciente na classe } D_2 \end{array}$$

sendo $f(\mathbf{x}/D_i)$, $i=1,2$, as funções de probabilidade nas duas classes.

Para uma variável dicotômica com

$$P(X=1/D_1)=p \qquad P(X=1/D_2)=q$$
$$P(X=0/D_1)=(1-p) \qquad P(X=0/D_2)=(1-q)$$

as regiões de alocação ótima, associadas à regra de classificação

I: aloca o paciente sempre em D_1

II { X=1: aloca em D_1
   { X=0: aloca em D_2

III $\begin{cases} X=1: \text{ aloca em } D_2 \\ X=0: \text{ aloca em } D_1 \end{cases}$

IV: aloca o paciente sempre em D_2

são apresentadas na figura a seguir

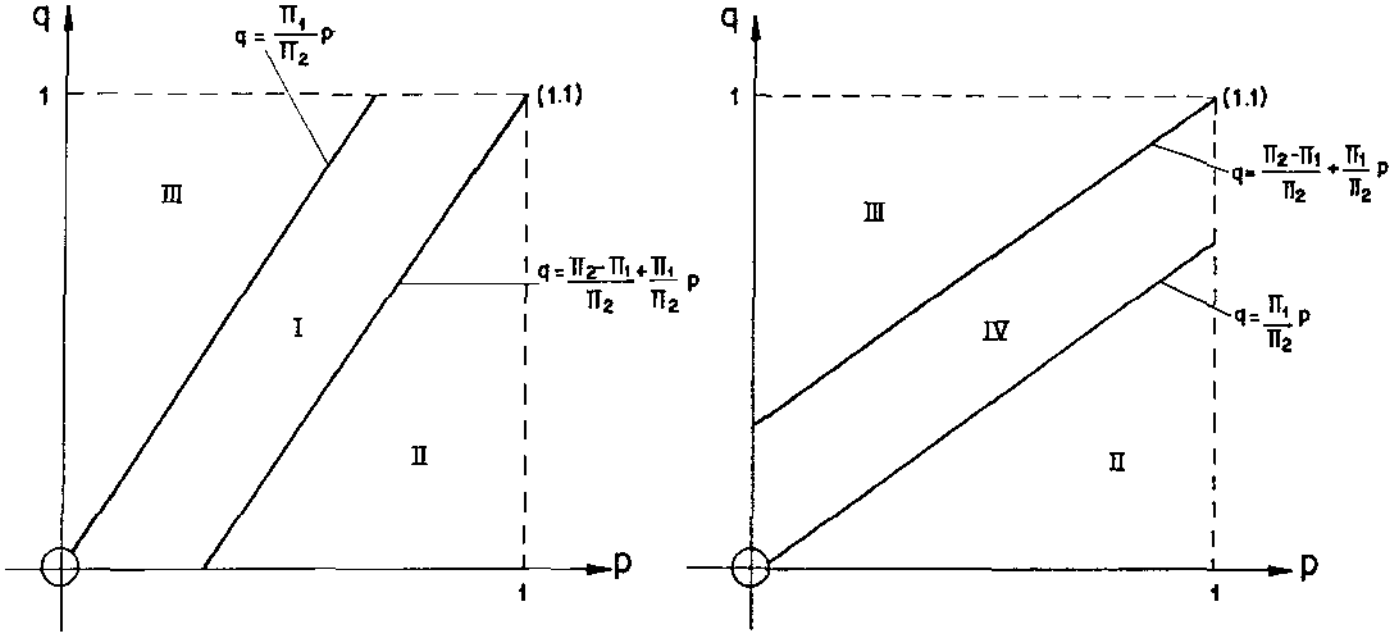


Figura 1: Regiões de alocação para o problema de classificação de dois grupos com probabilidades a priori difetentes e uma variável dicotônica(fig. 1a: $\pi_1 > \pi_2$, fig. 1b: $\pi_1 < \pi_2$).

Nas regiões I e IV a alocação é independente do valor observado, ou seja, a evidência fornecida por X não auxilia o processo de classificação.

O *erro ótimo* é definido como a probabilidade de má-classificação ao utilizarmos a regra de Bayes com os parâmetros conhecidos.

Para uma variável binária o *erro ótimo* é dado por:

$$F(X) = \min (\Pi_1 p , \Pi_2 q) + \min (\Pi_1 (1-p) , \Pi_2 (1-q)) \qquad (1)$$

A extensão(ou grau) das regiões apresentadas na figura 1 depende da diferença entre Π_1 e Π_2. Considerando $\Pi_1=\Pi_2$, as regiões são separadas pela reta p=q no quadrado unitário, como se percebe na figura abaixo. As paralelas a esta reta correspondem com taxas ótimas de erro iguais.

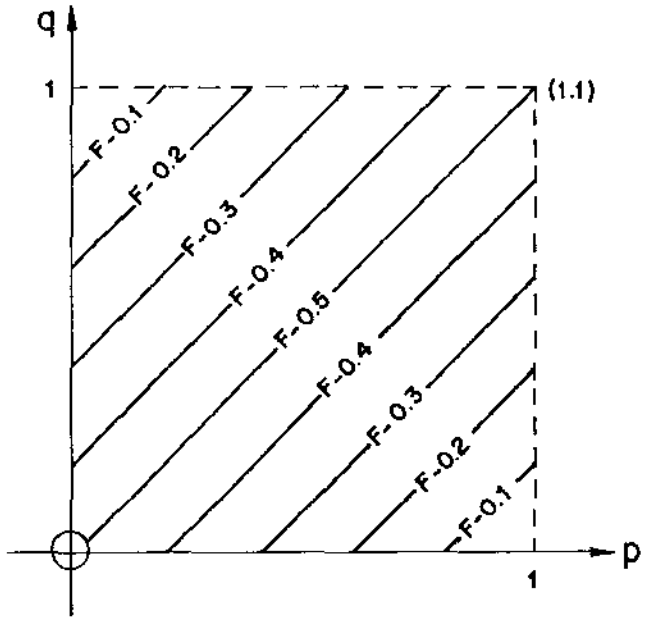


Figura 2: Taxa ótima de erro *F* para o problema de classificação em dois grupos com probabilidades a priori iguais($\pi_1=\pi_2$).

Aplicando a regra de Bayes para um par arbitrário de variáveis (X_i, X_j) chega-se a

$$F(X_i, X_j) = \sum_{l,m=0}^{1} \min (\Pi_1 p_{lm} , \Pi_2 q_{lm}) \qquad (2)$$

Considerando prioris iguais às classes D_1 e D_2, obtemos

$$F(X_i, X_j) = \frac{1}{2}\left[\, 1 - \frac{1}{2} \sum_{l,m=0}^{1} | p_{lm} - q_{lm} | \,\right]$$

Desde que $q_i \geq p_i \quad (i=1,\ldots,d)$, isto é, a probabilidade de presença dos indicadores é sempre maior no grupo D_2 (ou em D_1, desde que se redefina simbolicamente D_2 como o grupo de "doentes" e D_1 como os "sadios"), Haerting(1983) estabelece que a taxa ótima de erro do par (X_i, X_j) é dada por

$$F(X_i, X_j) = \frac{1}{2}\left[F(X_i) + F(X_j) - |F(X_i) - \delta(X_i, X_j)| + |F(X_j) - \delta(X_i, X_j)|\right] \tag{3}$$

$$\text{com} \quad \delta(X_i, X_j) = \frac{1}{2}(1 - q_i q_j + p_i p_j)$$

Assim, ao calcularmos a taxa de erro de um par de variáveis dicotômicas independentes não é suficiente considerar as taxas de erro das variáveis individuais mas também um termo $\delta(X_i, X_j)$ que depende de ambas as variáveis.

Além disso, o autor afirma que ao selecionarmos variáveis de uma seqüência ordenada de acordo com suas taxas de erro não é válido concluir que as taxas de erro dos correspondentes pares de variáveis tenham uma ordem correspondente.

Logo, se $F(X_1) < F(X_2) < F(X_3)$, por exemplo, a relação $F(X_1, X_2) < F(X_2, X_3)$ pode não ocorrer em todos os casos, como no exemplo abaixo:

$$\begin{array}{lll} X_1: & p_1=0.10 & q_1=0.90 \\ X_2: & p_2=0.05 & q_2=0.80 \\ X_3: & p_3=0.01 & q_3=0.71, \end{array}$$

de (1) teríamos

$$F(X_1)=0.10$$
$$F(X_2)=0.125$$
$$F(X_3)=0.15$$

Assumindo independência de todos os pares das três variáveis consideradas temos de (2) as seguintes taxas de erro

$$F(X_1,X_2)=0.0825$$
$$F(X_1,X_3)=0.069$$
$$F(X_2,X_3)=0.0575$$

Esta propriedade de não-monotonicidade na ordenação da seleção de pares de variáveis independentes dicotômicas é surpreendente em contraste com o comportamento de variáveis normais independentes na análise discriminante linear, onde o melhor sub-grupo de k variáveis é composto pelas k melhores variáveis individuais. Haerting afirma que no passado este fato foi negligenciado em todas as aplicações da regra de Bayes com variáveis independentes na elaboração de diagnóstico médico.

A observação acima foi feita por outros pesquisadores como Elashoff *et. al.*(1967) e depois por Toussaint(1971) e Cover(1974), mas segundo Haerting pode-se chegar à conclusões incorretas devido à fórmulas erradas como a apresentada por Elashoff.

Ao considerar a questão, sob as suposições

i) $F(X_1) < F(X_2) < F(X_3)$

ii) $q_i > p_i$ para $i=1,2,3$

iii) $q_1-p_1 > q_2-p_2 > q_3-p_3$

com $l_i=q_i-p_i$

$$h_i = \frac{1}{2}(1-p_i-q_i)$$

$$T_{ij} = |h_i| - |h_j|$$

para $i,j=1,2,3$

$i \neq j$

Toussaint afirma que uma condição suficiente para a situação citada ocorrer é dada por

$$|h_1| > |h_3|$$

e uma condição suficiente e necessária é dada por

$$T_{31} < \frac{1}{2}\left[\frac{I_1 - I_3}{I_2}\right](1 + 2|h_2|)$$

Ao considerarmos dependência entre as variáveis dicotômicas, introduzindo a covariância como parâmetro de interação para cada classe da doença, podemos escrever a distribuição de probabilidade de um par de variáveis (X_i, X_j) como

$$\begin{aligned} p_{11} &= p_i p_j + \rho & q_{11} &= q_i q_j + \tau \\ p_{10} &= p_i(1-p_j) - \rho & q_{10} &= q_i(1-q_j) - \tau \\ p_{01} &= (1-p_i)p_j - \rho & q_{01} &= (1-q_i)q_j - \tau \\ p_{00} &= (1-p_i)(1-p_j) + \rho & q_{00} &= (1-q_i)(1-q_j) + \tau \end{aligned}$$

sendo ρ e τ as covariâncias entre X_i e X_j nas classes D_1 e D_2, respectivamente, com os índices i, j omitidos para ρ e τ, onde

$$\rho \geq \max\{ -[p_i(1-p_i)p_j(1-p_j)]^{1/2}, -p_i p_j, p_i(1-p_j) - 1, (1-p_i)p_j - 1, -(1-p_i)(1-p_j) \}$$

e

$$\rho \leq \min\{ [p_i(1-p_i)p_j(1-p_j)]^{1/2}, 1 - p_i p_j, p_i(1-p_j), (1-p_i)p_j, 1 - (1-p_i)(1-p_j) \}$$

e o mesmo valendo para τ, com as correspondentes probabilidades q_i e q_j.

Assumindo ρ e τ conhecidas e usando

$$\Delta_{lm} = p_{lm} - q_{lm} \qquad l, m = 0, 1$$

obtemos de (3):

$$F(X_i, X_j) = \frac{1}{2}\left[1 - \frac{1}{2}\sum_{l,m}^{1} |\Delta_{lm}|\right] \qquad (4)$$

$$= \frac{1}{2}\left[1 - \frac{1}{2}|(1-p_i)(1-p_j) - (1-q_i)(1-q_j) + (\rho - \tau)|\right.$$
$$- \frac{1}{2}|(1-p_i)p_j - (1-q_j)q_j - (\rho - \tau)|$$
$$- \frac{1}{2}|p_i(1-p_j) - q_i(1-q_j) - (\rho - \tau)|$$
$$\left. - \frac{1}{2}|p_i p_j - q_i q_j + (\rho - \tau)|\right]$$

Assim, somente a diferença entre os parâmetros de covariância necessita ser conhecida ao calcularmos a taxa de erro de um par de variáveis dicotômicas. Se esta diferença é nula, as taxas de erro dos pares de variáveis dependentes se comportam como as variáveis independentes.

Covariâncias desiguais nas duas classes D_1 e D_2 resultam em alguns casos em um decréscimo, em outros em um acréscimo ou em não alteração nas taxas de erro, se comparadas com as correspondentes variáveis independentes. Haerting cita outro trabalho seu, onde se encontra uma análise detalhada sobre este comportamento, e o resultado devido a Cochran(1964) e Skarabis(1970), onde para variáveis normais com igual correlação nos dois grupos, correlações negativas sempre diminuem a taxa de erro enquanto correlações positivas somente diminuem a taxa se estiverem acima de um certo valor.

Assumindo-se independência geral entre as variáveis nos grupos e aplicando a "seleção passo a frente", a adição de uma nova variável nunca aumenta a taxa de erro. Entretanto, um método de busca deste tipo em geral não chega ao sub-grupo ótimo de variáveis de tamanho fixado.

- Mesmo sob a suposição de independência, pode acontecer que para tamanhos amostrais fixados a taxa atual de erro em regras de classificação estimadas decresça com o aumento do número de variáveis até um certo valor, aumentando novamente a partir daí.

A *taxa de erro atual*, definida como a probabilidade de má-classificação que resulta quando a regra de alocação é construída com estimativas dos parâmetros, é uma variável aleatória que Cochran e Hopkins(1961) mostraram ter esperança sempre maior ou igual ao erro ótimo.

Vacek(1985) estudou o efeito da dependência dos testes clínicos na taxa de erro e estimadores das prevalências da doença para situações em que ambos os testes tem taxas de erro desconhecidas. O trabalho, associado a testes aplicados simultaneamente a duas populações com prevalências diferentes da doença, encontrou resultados similares a Thibodeau(1981). Este investigou o efeito de uma correlação positiva entre testes diagnósticos, quando a taxa de erro de um novo teste é estimada por comparação com um teste referência que tem taxa de erro já conhecida.

Thibodeau encontrou limites inferiores para a sensibilidade e especificidade do novo teste, demonstrando que estes limites podem ser bem amplos, particularmente quando a eficiência do teste referência é baixa. Assim, a suposição de independência pode resultar em um sub-estimação da taxa de erro de um novo teste, se ele for positivamente correlacionado com o teste referência.

Além disso, Vacek encontrou que os estimadores das taxas de prevalência nas duas populações podem ser positivamente ou negativamente viciados, dependendo da magnitude relativa das duas covariâncias condicionais e o valor do parâmetro de prevalência.

Sob outro contexto, ao analisar um exemplo, Pereira e Pericchi(1985) concluem que se a independência condicional é utilizada no estudo do peso da evidência, ocorre uma sub-estimação da dispersão das caselas e uma super-estimação do poder de sensibilidade e especificidade do efeito conjunto de dois testes clínicos.

3. O CRITÉRIO DE KOKOLAKIS PARA SELEÇÃO DE VARIÁVEIS

3.1 Introdução

Considerando ainda o problema de classificar um paciente a um dos grupos de $D = \{D_1, D_2\}$, com $\pi_1 = P(X \in D_1)$ e $\pi_2 = 1 - \pi_1$ conhecidas, duas amostras de tamanhos m_1 e m_2 de um vetor d-dimensional $X = (X_1, \ldots, X_d)'$ de indicadores médicos são retiradas dos grupos D_1 e D_2, respectivamente.

Sejam θ_i $(i=1,\ldots,n=2^d)$ as chances das caselas das 2^d possíveis seqüências binárias de D_1 e ψ_i $(i=1,..,n)$ aqueles de D_2.

As distribuições a priori conjugadas para estes parâmetros são Dirichlet ou misturas de prioris Dirichlet. Contudo, Brown (1976) mostrou que estas prioris são irrealistas. Em primeiro lugar, a média a posteriori da chance de qualquer casela depende somente na frequência da correspondente casela nas amostras piloto, tornando difícil a classificação quando o número $n=2^d$ de caselas é próximo do tamanho amostral m.

Além disso, a priori conjugada não permite correlação positiva entre as probabilidades das caselas, quando seria razoável esperar que as chances associadas à seqüências similares (como d-1 coincidências nos d indicadores médicos) fossem correlacionadas positivamente. Ao considerar dois pacientes que diferem somente em um sintoma, como $\mathbf{x}_{(A)}=(1,0,1,1,1)$ e $\mathbf{x}_{(B)}=(1,0,1,1,0)$, Lindley(1978) julga que as duas chances de que pertençam à classe D_1 devam ser correlacionadas. Em linguagem médica, pacientes com 4 sintomas comuns em 5 são "similares", no sentido de que se um deles é da classe D_i, é provável que o outro também o seja. Este tipo de correlação parece essencial ao autor para evitar a dificuldade de que mesmo com grande tamanho amostral dois pacientes não são iguais, de tal forma que $\mathbf{x}$ é diferente para cada paciente de D_1 e D_2. Além disso, Lindley cita o estudo de Hughes(1968) que calculou a probabilidade de classificação correta p_n, com classificação baseada na maior probabilidade a

posteriori e ignorando qualquer consideração sobre utilidade. Ao assumir que $\underset{\sim}{\theta}$ (e independentemente Ψ) é uniformemente distribuída sobre o n-simplex $\theta_i \geq 0$, $\sum\theta_i = 1$, Hughes encontrou

$$p_n = \frac{1}{2} \sum_i E \max\{\theta_i, \psi_i\} = \frac{3n - 2}{2(2n - 1)}$$

$$\text{com} \quad \lim_{n \to \infty} p_n = \frac{3}{4}$$

(assumindo a disponibilidade de uma amostra piloto de tamanho infinito). Entretanto, supondo $n=2^s$ e que o refinamento é realizado pela adição de uma variável extra X_{s+1} tal que a probabilidade de $X_{s+1}=1$ dado qualquer conjunto $X_1,\ldots,X_s$ é uniforme no intervalo unitário, todas elas sendo independentes, Lindley encontra $\lim p_{(s)}=1$ ($s \to \infty$), com $p_{(s)}$ sempre excedendo p_n ($n=2^s$).

Sob tal enfoque, na classificação com duas variáveis binárias, por exemplo, onde as chances de que $X_1=1$, de que $X_2=1$ dado $X_1=0$ e dado $X_1=1$, são todas independentes e uniformes no intervalo unitário, Lindley encontra que a chance λ de que $X_2=1$ tem a densidade marginal

$$-2 [\lambda \log \lambda + (1-\lambda) \log (1-\lambda)]$$

que é mais concentrada em torno de $\lambda=1/2$ do que a distribuição uniforme. Consequentemente, a ordem de X_1, X_2 é relevante e atribuir esta distribuição significa que estamos mais incertos sobre X_1 do que sobre X_2. Embora a distribuição de X_i ($i > 2$) não seja conhecida sua variância é igualmente menor, o que não seria razoável na prática e certamente também não com sintomas médicos.

Assim, é requerido que todas as chances de $X_i=1$ ($i=1,\ldots,d$) tenham a mesma distribuição marginal e similarmente pares de variáveis, trios e assim por diante, ou seja, uma forma de permutabilidade parcial nestas chances.

Com d=2 temos as seis chances donominadas doravante de chances condicionais, contrastando com os θ's:

$u = P(X_1=1)$ $\qquad$ $r = P(X_2=1)$

$v = P(X_2=1/X_1=1)$ $\qquad$ $s = P(X_1=1/X_2=1)$

$w = P(X_2=1/X_1=0)$ $\qquad$ $t = P(X_1=1/X_2=0)$

sendo requerido que os vetores aleatórios (u,v,w) e (r,s,t) sejam identicamente distribuídos.

Assim, a ordem em que os dois testes seriam considerados não deve afetar a especificação a priori para as chances condicionais, ou seja, esta priori é requerida invariante sob permutação dos testes.

Good (1965, apêndice E) considera este problema sob enfoque adicional requerendo que todas estas chances condicionais sejam independentes com distribuições idênticas e simétricas em torno de 1/2. Entretanto, conclui que não existe uma distribuição contínua satisfazendo estas exigências.

Kokolakis(1983) caracteriza uma família de prioris que são invariantes sob permutações dos testes.

Seja θ_i (i=1,...,n) as chances das caselas das seqüências d-dimensionais de 1's (presença) e 0's (ausência) em ordem lexicográfica, com 1 antes de 0. Temos

$$\theta_1 = P(X_1=1, X_2=1, \ldots, X_d=1/D_1)$$

$$\theta_2 = P(X_1=1, X_2=1, \ldots, X_{d-1}=1, X_d=0/D_1)$$

$$\vdots$$

$$\theta_n = P(X_1=0, X_2=0, \ldots, X_d=0/D_1)$$

Consideremos também as seguintes probabilidades condicionais, restritas somente a $D=D_1$:

$$u_{11} = P(X_1=1)$$

$$u_{21} = P(X_2=1/X_1=1)$$

$$u_{22} = P(X_2=1/X_1=0)$$

$$u_{31} = P(X_3=1/X_1=1, X_2=1)$$

$$u_{32} = P(X_3=1/X_1=1, X_2=0)$$

$$u_{33} = P(X_3=1/X_1=0, X_2=1)$$
$$u_{34} = P(X_3=1/X_1=0, X_2=0)$$
$$\vdots$$

$$u_{ij} = P(X_i=1/X_1=x_1,\ldots,X_{i-1}=x_{i-1}) \qquad i=1,\ldots,d \quad j=1,\ldots,2^{i-1}$$

Assim,

$$\theta_1 = u_{11}\ u_{21}\ \ldots\ u_{d-1,1}\ u_{d,1}$$
$$\theta_2 = u_{11}\ u_{21}\ \ldots\ u_{d-1,1}\ (1 - u_{d,1})$$
$$\vdots$$

$$\theta_n = (1 - u_{11})(1 - u_{21})\ldots(1 - u_{d-1,2^{d-2}})(1 - u_{d,2^{d-1}})$$

ou seja,

$$\theta = T(u)$$

onde T é uma transformação 1-1, isto é

$$U = T^{-1}(\theta)$$

Assim, pode-se trabalhar de forma equivalente com θ ou o correspondente u.

Para obter uma distribuição para θ (ou u) invariante sob todas as permutações dos testes, precisamos somente de invariância sobre P_k (k=2,...,d), onde P_k transpõe os testes (k-1) e k.

Lema: A permutação P_k nos testes implica na seguinte permutação Q_k nas chances das caselas

$$\theta_{(4m+1)2^{d-k}+l} \longleftrightarrow \theta_{(4m+2)2^{d-k}+l} \qquad m=0,1,..,2^{k-2}-1 \quad l=1,\ldots,2^{d-k}$$

permanecendo inalterados os demais θ's.

Exemplo:

X_1	X_2	X_3	θ_i
1	1	1	θ_1
1	1	0	θ_2
1	0	1	θ_3
1	0	0	θ_4
0	1	1	θ_5
0	1	0	θ_6
0	0	1	θ_7
0	0	0	θ_8

$\xrightarrow{P_2}$

X_2	X_1	X_3	θ_i
1	1	1	θ_1
1	1	0	θ_2
1	0	1	θ_5
1	0	0	θ_6
0	1	1	θ_3
0	1	0	θ_4
0	0	1	θ_7
0	0	0	θ_8

Logo,

$$P_2 \longrightarrow \begin{matrix} m=0 \\ I=1,2 \end{matrix}$$

$$\theta_3 \longleftrightarrow \theta_5$$
$$\theta_4 \longleftrightarrow \theta_6$$

Os elementos do grupo de permutações gerado por Q_k $(k=2,\ldots,d)$ serão denotados por Q_i^* $(i=1,\ldots,d!)$.

Corolário: Permutabilidade parcial entre os θ's correspondentes à seqüências com o mesmo número de 1's implica uma distribuição para $\underset{\sim}{\theta}$ (ou **U**) invariante sob permutações dos testes.

Isto ocorre porque qualquer permutação dos testes implica uma troca entre os θ's com o mesmo número de 1's.

Em particular, teremos que toda distribuição Dirichlet simétrica induzirá distribuições invariantes sob permutações dos testes para as chances condicionais.

3.2 Mistura generalizada de prioris Dirichlet

Suponha que as chances condicionais u_{ij} $(i=1,\ldots,d,\ j=1,\ldots,2^{i-1})$ são quantidades aleatórias independentes com densidade *beta*, ou seja,

$$p(u_{ij}) = \frac{\Gamma(\alpha_{2j-1}^{(i)} + \alpha_{2j}^{(i)})}{\Gamma(\alpha_{2j-1}^{(i)})\Gamma(\alpha_{2j}^{(i)})}\ u_{ij}^{\alpha_{2j-1}^{(i)}-1}\ (1-u_{ij})^{\alpha_{2j}^{(i)}-1}, \qquad (5)$$

$$u_{ij} \in (0,1), \quad \text{com} \quad \alpha_{2j-1}^{(i)},\ \alpha_{2j}^{(i)} > 0.$$

Tem-se que a função densidade de $\underset{\sim}{\theta}$ é dada por

$$g(\underset{\sim}{\theta}) = C(\underset{\sim}{\alpha}^{(1)},\ldots,\underset{\sim}{\alpha}^{(d)}) \left[\prod_{i=1}^{n} \theta_i^{\alpha_i^{(d)}-1}\right]$$

$$\left[\left(\sum_{i=1}^{n/2} \theta_i\right)^{\alpha_1^{(2)}+\alpha_2^{(2)}-\alpha_1^{(1)}} \left(\sum_{i=(n/2)+1}^{n} \theta_i\right)^{\alpha_3^{(2)}+\alpha_4^{(2)}-\alpha_2^{(1)}} \cdots \right. \qquad (6)$$

$$\left(\sum_{i=1}^{2} \theta_i\right)^{\alpha_1^{(d)}+\alpha_2^{(d)}-\alpha_1^{(d-1)}} \cdots$$

$$\left.\left(\sum_{i=n-1}^{n} \theta_i\right)^{\alpha_{n-1}^{(d)}+\alpha_n^{(d)}-\alpha_{n/2}^{(d-1)}}\right]^{-1}, \quad \underset{\sim}{\theta} \in S^{(n)}$$

onde $S^{(n)}$ indica o simplex n-dimensional e a constante normalizadora é dada por

$$C(\underset{\sim}{\alpha}^{(1)},\ldots,\underset{\sim}{\alpha}^{(d)}) = \prod_{i=1}^{d}\prod_{j=1}^{2^{i-1}} \frac{\Gamma(\alpha_{2j-1}^{(i)} + \alpha_{2j}^{(i)})}{\Gamma(\alpha_{2j-1}^{(i)})\Gamma(\alpha_{2j}^{(i)})}$$

Se os parâmetros α's satisfazem

$$\alpha_{2j-1}^{(i)} + \alpha_{2j}^{(i)} = \alpha_{j}^{(i-1)} \qquad \begin{matrix} i=2,\ldots,d \\ j=1,\ldots,2^{i-1} \end{matrix}$$

ou seja,

$$\alpha_1^{(1)} + \alpha_2^{(1)} = \alpha_1^{(0)}$$

$$\alpha_1^{(2)} + \alpha_2^{(2)} = \alpha_1^{(1)}$$

$$\alpha_3^{(2)} + \alpha_4^{(2)} = \alpha_2^{(1)}$$

$$\vdots$$

temos de (6) a distribuição de Dirichlet com parâmetros $\underset{\sim}{\alpha}^{(d)}$.

Como $g(\underset{\sim}{\theta})$ não satisfaz a permutabilidade para os θ's especificados para invariância, p(U) não será invariante sob permutação dos testes.

Entretanto, tomando a mistura

$$p(\underset{\sim}{\theta}) = \frac{1}{d!} \sum_{i=1}^{d!} g\{Q_i^{*}(\underset{\sim}{\theta})\}, \qquad \underset{\sim}{\theta} \in S^{(n)} \tag{7}$$

a função densidade induzida de U será invariante.

Segundo o autor, este meio de produzir uma priori invariante parece ser complicado em demasia, mas seria de fato a única maneira ao se utilizar núcleos da forma como (6).

Além disso, estes núcleos são tratáveis matematicamente e implicam em correlação positiva entre as chances das caselas correspondentes à seqüências que tenham um número relativamente grande de elementos em comum.

Apenas por razões de simplicidade assume-se que os hiperparâmetros α's em (5) são todos iguais a uma constante positiva α.

Assim, $g(\underset{\sim}{\theta})$ dado em (6) se tornaria

$$g(\theta) = C(\alpha) \frac{\prod_{i=1}^{n} \theta_i^{\alpha-1}}{\left[\left(\sum_{i=1}^{n/2} \theta_i\right)\left(\sum_{i=(n/2)+1}^{n} \theta_i\right)\cdots\left(\sum_{i=1}^{2} \theta_i\right)\cdots\left(\sum_{i=n-1}^{n} \theta_i\right)\right]^{\alpha}},$$

$$\underset{\sim}{\theta} \in S^{(n)} \tag{8}$$

de forma que os u's serão independentes com distribuição beta de parâmetro α, o que não ocorrerá para as distribuições dos u's derivados de (7).

Exemplo: Utilizando duas variáveis (d=2) para classificar populações, com

$P(X_1, X_2/D_1)$

X_1 \ X_2	1	0
1	θ_1	θ_2
0	θ_3	θ_4

$P(X_1, X_2/D_2)$

X_1 \ X_2	1	0
1	ψ_1	ψ_2
0	ψ_3	ψ_4

temos de (7) e (8)

$$p(\theta_1,\theta_2,\theta_3,\theta_4) = \frac{1}{2}\left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)^2}\right]^3$$

$$\left[\frac{1}{[(\theta_1+\theta_2)(\theta_3+\theta_4)]^{\alpha}} + \frac{1}{[(\theta_1+\theta_3)(\theta_2+\theta_4)]^{\alpha}}\right] \prod_{i=1}^{4} \theta_i^{\alpha-1} \tag{9}$$

Trabalhando com as chances condicionais já definidas

$u = P(X_1=1) = \theta_1+\theta_2$ $\qquad$ $r = P(X_2=1) = \theta_1+\theta_3$

$$v = P(X_2=1/X_1=1) = \frac{\theta_1}{\theta_1+\theta_2} \qquad s = P(X_1=1/X_2=1) = \frac{\theta_1}{\theta_1+\theta_3}$$

$$w = P(X_2=1/X_1=0) = \frac{\theta_3}{\theta_3+\theta_4} \qquad t = P(X_1=1/X_2=0) = \frac{\theta_2}{\theta_2+\theta_4}$$

o autor encontra que (u,v,w) e (r,s,t) são identicamente distribuídos e as chances condicionais distribuídas simetricamente em torno de $\frac{1}{2}$.

Além disso, de (9) Kokolakis (1983) encontra que os θ's são individualmente distribuídos como o produto de duas variáveis aleatórias independentes com distribuição beta simétrica de parâmetro α.

A distribuição preditiva de $X' = (X_1, X_2)' \in \{(1,1),(1,0),(0,1),(0,0)\}$ é dada por

$$p(X/r) = \frac{R_{X_1X_2} + \alpha}{m + 2\alpha} \left[\frac{R_{X_1} + \alpha}{R_{X_1} + 2\alpha} W_1 + \frac{R_{X_2} + \alpha}{R_{X_2} + 2\alpha} W_2 \right]$$

com

$$W_1 = \left[1 + \frac{\Gamma(r_1+r_3+\alpha)\Gamma(r_2+r_4+\alpha)\Gamma(r_1+r_2+2\alpha)\Gamma(r_3+r_4+2\alpha)}{\Gamma(r_1+r_2+\alpha)\Gamma(r_3+r_4+\alpha)\Gamma(r_1+r_3+2\alpha)\Gamma(r_2+r_4+2\alpha)} \right]^{-1} = 1 - W_2$$

$R_{X_1 \ldots X_k}$ $(k=1,\ldots,d)$ a frequência do vetor $(X_1,\ldots,X_k)$ na amostra piloto e $r = (r_1,r_2,r_3,r_4)'$ a estatística suficiente para $\underset{\sim}{\theta} = (\theta_1,\theta_2,\theta_3,\theta_4)'$, $m = \sum_{i=1} r_i$.

Quando $\alpha = 1$, o peso W_1 é aproximadamente inversamente proporcional à estimativa amostral da variância [igual a $\frac{(r_1+r_2)(r_3+r_4)}{m^2}$] do primeiro teste, com caso similar para W_2.

Com d>2 em (7) e $g(\underset{\sim}{\theta})$ dada por (8) todas as chances condicionais u_{ij} são simetricamente distribuídas em torno de 1/2, todas as chances condicionais da mesma geração, por exemplo k,

$$u_{kj} \quad (j=1,\ldots,2^{k-1})$$

são identicamente distribuídas, e as chances das caselas θ_i $(i=1,\ldots,n=2^d)$ são distribuídas como o produto de d variáveis aleatórias independentes simétricas beta de parâmetro α.

Além disso, para duas chances θ_i e θ_j com l $(0 \leq l < 1)$ elementos em comum no vetor de realizações binárias, tem-se

$$C_l = \text{Cov}\,(\theta_i,\theta_j) = \frac{1}{4^d}\left[\frac{2\alpha}{1+2\alpha}\sum_{t=0}^{l}\frac{\Gamma(l+1)\Gamma(d-l+1)}{\Gamma(l-t+1)\Gamma(d-l+t+1)}\frac{1}{(1+2\alpha)^t} - 1\right]$$

função com comportamento monotonicamente crescente em relação a l.

Dois pacientes com um quadro clínico de indicadores muito parecidos, por exemplo l = d-1 coincidências entre os vetores de realizações binárias, apresentam a covariância entre as correspondentes chances das caselas

$$C_{d-1} = \frac{1}{4^d}\left[\frac{2\alpha}{d}\left[\left[\frac{2(1+\alpha)}{1+2\alpha}\right]^d - 1\right] - 1\right]$$

com $C_{d-1}>0$ desde que $\left[\frac{2(1+\alpha)}{1+2\alpha}\right]^d > 1 + \frac{d}{2\alpha}$,

o que é satisfeito para um número d de indicadores suficientemente grande.

A distribuição preditiva de $X' = (X_1,\ldots,X_d)'$ dado r é dada por

$$p(X/r) = \frac{R_X+\alpha}{m+2\alpha}\sum_{i=1}^{d!} V_i(X)\, W_i \qquad (10)$$

onde

$$V_i(X) = \frac{R_{X_{i_1}\ldots X_{i_{d-1}}} + \alpha}{R_{X_{i_1}\ldots X_{i_{d-1}}} + 2\alpha} \cdots \frac{R_{X_{i_1}} + \alpha}{R_{X_{i_1}} + 2\alpha} \qquad (i=1,\ldots,d!)$$

sendo $(X_{i_1},\ldots,X_{i_d})$ a permutação de $(X_1,\ldots,X_d)$ que corresponde à permutação Q_i^* dos θ's, $R_{X_{i_1}\ldots X_{i_k}}$ $(k=1,\ldots d)$ a frequência do vetor $(X_{i_1},\ldots,X_{i_n})$ nos dados amostrais

e

$$W_i = \left[\sum_{j=1}^{d!} \frac{C(\alpha;Q_i^*(r))}{C(\alpha;Q_j^*(r))} \right]^{-1}, \quad i=1,\ldots,d!$$

onde

$$C(\alpha;r) = \frac{\Gamma\left(\sum_{i=1}^{n} r_i + 2\alpha\right)}{\Gamma\left(\sum_{i=1}^{n/2} r_i + \alpha\right)\Gamma\left(\sum_{i=(n/2)+1}^{n} r_i + \alpha\right)}$$

$$\cdot \frac{\Gamma\left(\sum_{i=1}^{n/2} r_i + 2\alpha\right)\ \Gamma\left(\sum_{i=(n/2)+1}^{n} r_i + 2\alpha\right)}{\Gamma\left(\sum_{i=1}^{n/4} r_i + \alpha\right)\Gamma\left(\sum_{i=(n/4)+1}^{n/2} r_i + \alpha\right)\Gamma\left(\sum_{i=(n/2)+1}^{3n/4} r_i + \alpha\right)\Gamma\left(\sum_{i=(3n/4)+1}^{n} r_i + \alpha\right)}$$

$$\cdots \frac{\Gamma(r_1+r_2+2\alpha)\ \ldots\ \Gamma(r_{n-1}+r_n+2\alpha)}{\Gamma(r_1+\alpha)\Gamma(r_2+\alpha)\ \ldots\ \Gamma(r_n+\alpha)}$$

3.3 Um modelo para a seleção de variáveis

Suponha

$$\alpha_{2j-1}^{(i)} = \alpha_{2j}^{(i)} = \alpha > 0 \qquad \begin{matrix} i=1,\ldots,k \\ j=1,\ldots 2^{i-1} \end{matrix}$$

$$\alpha_{2j-1}^{(i)} = \alpha_{2j}^{(i)} = \frac{1}{2}\,\alpha_{j}^{(i-1)}, \qquad \begin{matrix} i=k+1,\ldots,d \\ j=1,\ldots,2^{i-1} \end{matrix}$$

para a densidade dada em (6).

Estas suposições são equivalentes a assumir igualdade dos "pesos" $\gamma(\theta_i^{(d)})$ das seguintes chances de caselas:

$$\alpha_1^{(1)} = \alpha_2^{(1)} = \alpha$$

isto é,

$$\gamma(\theta_1^{(1)}) = \gamma(\theta_2^{(1)})$$

$$\gamma[P(X_1=1)] = \gamma[P(X_1=0)] = \alpha$$

$$\alpha_1^{(2)} = \alpha_2^{(2)} = \alpha$$

$$\gamma[P(X_1=1,X_2=1)] = \gamma[P(X_1=1,X_2=0)] = \alpha$$

$$\alpha_3^{(2)} = \alpha_4^{(2)} = \alpha$$

$$\gamma[P(X_1=0,X_2=1)] = \gamma[P(X_1=0,X_2=0)] = \alpha$$

$$\alpha_1^{(3)} = \alpha_2^{(3)} = \alpha$$

$$\gamma[P(X_1=1,X_2=1,X_3=1)] = \gamma[P(X_1=1,X_2=1,X_3=0)] = \alpha$$

$$\vdots$$

$$\alpha_7^{(3)} = \alpha_8^{(3)} = \alpha$$

$$\gamma[P(X_1=0, X_2=0, X_3=1)] = \gamma[P(X_1=0, X_2=0, X_3=0)] = \alpha$$

$$\vdots$$

$$\alpha_1^{(k)} = \alpha_2^{(k)} = \alpha$$

$$\gamma[P(X_1=1, \ldots, X_k=1)] = \gamma[P(X_1=0, \ldots, X_{k-1}=1, X_k=0)] = \alpha$$

$$\vdots$$

$$\alpha_{2^{k-1}}^{(k)} = \alpha_{2^k}^{(k)} = \alpha$$

$$\gamma[P(X_1=0, \ldots, X_{k-1}=0, X_k=1)] = \gamma[P(X_1=0, \ldots, X_k=0)] = \alpha$$

$$\alpha_1^{(k+1)} = \alpha_2^{(k+1)} = \frac{1}{2}\,\alpha_1^{(k)}$$

$$\gamma[P(X_1=1, \ldots, X_{k+1}=1)] = \gamma[P(X_1=1, \ldots, X_k=1, X_{k+1}=0)]$$

$$= \frac{1}{2}\,\gamma[P(X_1=1, \ldots, X_k=1)] = \frac{\alpha}{2}$$

$$\alpha_3^{(k+1)} = \alpha_4^{(k+1)} = \frac{1}{2}\,\alpha_2^{(k)}$$

$$\gamma[P(X_1=1, \ldots, X_k=0, X_{k+1}=1)] = \gamma[P(X_1=1, \ldots, X_k=0, X_{k+1}=0)]$$

$$= \frac{1}{2}\,\gamma[P(X_1=1, \ldots, X_{k-1}=1, X_k=0)] = \frac{\alpha}{2}$$

$$\vdots$$

$$\alpha_1^{(k+2)} = \alpha_2^{(k+2)} = \frac{1}{2}\,\alpha_1^{(k+1)}$$

$$\gamma[P(X_1=1, \ldots, X_{k+2}=1)] = \gamma[P(X_1=1, \ldots, X_{k+1}=1, X_{k+2}=0)]$$

$$= \gamma[P(X_1=1, \ldots, X_{k+1}=1)] = \frac{\alpha}{4}$$

$$\vdots$$

$$\alpha^{(k+2)}_{2^{k+2}-1} = \alpha^{(k+2)}_{2^{k+2}} = \frac{1}{2}\,\alpha^{(k+1)}_{2^{k+1}} = \frac{\alpha}{4}$$

$$\vdots$$

$$\alpha^{(d)}_1 = \alpha^{(d)}_2 = \frac{1}{2}\,\alpha^{(d-1)}_1$$

$$\gamma[P(X_1=1,\ldots,X_d=1)] = \gamma[P(X_1=1,\ldots,X_{d-1}=1,X_d=0)]$$

$$= \frac{1}{2}\,\gamma[P(X_1=1,\ldots,X_{d-1}=1)] = \frac{\alpha}{2^{d-k}}$$

$$\vdots$$

$$\alpha^{(d)}_{2^d-1} = \alpha^{(d)}_{2^d} = \frac{1}{2}\,\alpha^{(d-1)}_{2^{d-1}} = \frac{\alpha}{2^{d-k}}$$

Assim, o modelo de Kokolakis atribui pesos iguais às probabilidades das caselas das seguintes probabilidades

$$P(X_1=1),\ P(X_1=0),\ P(X_i=x_i,X_j=x_j),\ P(X_i=x_i,X_j=x_j,X_t=x_t),\ldots$$

para quaisquer variáveis entre as k fixadas como sub-grupo ótimo, privilegiando as probabilidades destas caselas.

Para as demais, acompanhando as k melhores, teriamos pesos divididos simetricamente (uniformemente) e cada vez menores a medida que aumenta o número destas variáveis "desnecessárias" no modelo de classificação.

Enquanto isso, para o vetor de parâmetros $\underset{\sim}{\alpha}^{(d)} = (\alpha^{(d)}_1,\ldots,\alpha^{(d)}_n)'$ da densidade de Dirichlet, ao atribuirmos

$$\alpha^{(1)}_1 = \alpha^{(1)}_2 = \alpha > 0$$

isto é,

$$\gamma[P(X_1=1)] = \gamma[P(X_1=0)] = \alpha$$

e

$$\alpha_{2j-1}^{(i)} = \alpha_{2j}^{(i)} = \frac{1}{2}\,\alpha_j^{(i-1)} \qquad \begin{matrix} i=2,\ldots d \\ j=1,\ldots,2^{i-1} \end{matrix}$$

teremos a densidade simétrica da Dirichlet com parâmetro $\frac{2\alpha}{n}$.

Então, sob a suposição dada no início desta seção, temos

$$g(\theta) \propto h(\theta) \equiv \frac{\prod_{i=1}^{n} \theta_i^{(2\alpha/v)-1}}{D(\theta)}, \quad \theta \in S^{(n)} \tag{11}$$

onde

$$D(\theta) = \left[\left(\sum_{i=1}^{n/2}\theta_i\right)\left(\sum_{i=(n/2)+1}^{n}\theta_i\right)\right]\left[\left(\sum_{i=1}^{n/4}\theta_i\right)\cdots\left(\sum_{i=(3n/4)+1}^{n}\theta_i\right)\right]\cdots$$

$$\cdots\left[\left(\sum_{i=1}^{v}\theta_i\right)\cdots\left(\sum_{i=n-v+1}^{n}\theta_i\right)\right]^{\alpha}, \qquad v = 2^{d-k+1}.$$

A posteriori, dado a estatística suficiente $\mathbf{r} = (r_1,\ldots,r_n)'$, $\underset{\sim}{\theta}$ terá densidade

$$p(\underset{\sim}{\theta}/\mathbf{r}) = C(\alpha;\ \mathbf{r})\ \frac{\prod_{i=1}^{n}\theta_i^{r_i+(\alpha/v)-1}}{D(\theta)}, \qquad \theta \in S^{(n)}$$

onde

$$C(\alpha;\ \mathbf{r}) = \frac{\Gamma\left(\sum_{i=1}^{n} r_i + 2\alpha\right)}{\prod_{i=1}^{n}\Gamma(r_i + 2\alpha/n)}\left[\frac{\Gamma\left(\sum_{i=1}^{n/2} r_i + 2\alpha\right)\Gamma\left(\sum_{i=(n/2)+1}^{n} r_i + 2\alpha\right)}{\Gamma\left(\sum_{i=1}^{n/2} r_i + \alpha\right)\Gamma\left(\sum_{i=(n/2)+1}^{n} r_i + \alpha\right)}\cdots\right.$$

$$\left.\cdots\frac{\Gamma\left(\sum_{i=1}^{v} r_i + 2\alpha\right)\cdots\Gamma\left(\sum_{i=n-v+1}^{n} r_i + 2\alpha\right)}{\Gamma\left(\sum_{i=1}^{v} r_i + a\right)\cdots\Gamma\left(\sum_{i=n-v+1}^{n} r_i + 2\alpha\right)}\right]$$

Kokolakis (1985) mostrou que a chance de X_i ser igual a 1 é *a priori* simetricamente distribuída em torno de 1/2, com variância

$$\text{Var } [P(X_i=1)] = \begin{cases} \dfrac{1}{4(1+2\alpha)} \left[\dfrac{1+\alpha}{1+2\alpha}\right]^{i-1}, & i=1,\ldots,k \\ \dfrac{1}{4(1+2\alpha)} \left[\dfrac{1+\alpha}{1+2\alpha}\right]^{k-1}, & i=k+1,\ldots,d \end{cases}$$

Para d=2, por exemplo, isto pode ser encontrado utilizando

$$E(\theta_i^{\delta}) = \left[\frac{\Gamma(2\alpha)}{\Gamma(\alpha)} \frac{\Gamma(\alpha+\delta)}{\Gamma(2\alpha+\delta)}\right]^2 \qquad \begin{matrix} i=1,\ldots,4 \\ \delta=0,1,2,\ldots \end{matrix}$$

obtida de (9) e o fato de que os θ's são distribuídos individualmente como o produto de duas distribuições betas simétricas e independentes de parâmetro α.

Assim, por exemplo, o resultado pode ser achado para

$$u = P(X_i=1) = \theta_1 + \theta_2$$

A variância encontrada acima é uma função decrescente de i (i=1,...,k). Logo, o (i + 1)-ésimo teste, correspondendo à variável X_{i+1} é a priori esperado ser menos útil que o i-ésimo teste (i=1,...,k).

Isto parece intuitivo pois cada particular configuração do vetor $(X_1, X_2, \ldots, X_{i-1})$ deveria influenciar o valor de $P(X_i=1/X_1,\ldots,X_{i-1})$. A presença de muitos sintomas, por exemplo, deveria estar acompanhada por uma alta probabilidade para o ocorrência do i-ésimo sintoma, com raciocínio análogo à ausência da maioria deles.

Assim,

$P(X_i=1/X_1=1, X_2=1, \ldots, X_{i-2}=1, X_{i-1}=1)$ deveria ser alta

$P(X_i=1/X_1=0, X_2=0, \ldots, X_{i-1}=0)$ deveria ser baixa

logo, $P(X_1=1)$ deveria ser uma quantidade com muita oscilação, tornando $V[P(X_i=1)]$ alta ou pelo menos não-decrescente em função de i (ordem do teste).

Como $V[P(X_i=1)]$ é decrescente isto indica que o i-ésimo teste ($i=1,\ldots,k$) tende a apresentar o sintoma com uma probabilidade estável, independente dos sintomas anteriores já observados. Logo, para i suficientemente grande, o teste seria praticamente inútil em relação à contribuição que poderia dar para a separação dos grupos (diagnóstico do paciente no grupo sadio ou no grupo doente). Assim, os testes anteriores já estariam fornecendo a informação relevante para classificação.

Além disso, a priori $g(\theta)$ dada em (11) é invariante sob permutações dos últimos (d-k) testes mas não em relação aos primeiros k testes.

Contudo, tomando novamente uma mistura (como em (7), dada por Kokolakis (1983), atinge-se invariância com

$$p(\theta) \propto \frac{1}{N} \sum_{i=1}^{N} h\{Q_i(\theta)\} , \qquad \theta \in S^{(n)}$$

$$N = \frac{\Gamma(d+1)}{\Gamma(d-k+1)} = (d)_k$$

onde $Q_i(\theta)$ é uma permutação dos θ's correspondendo à ordenação $(i_1,\ldots,i_k)$ de k testes, do conjunto $\{1,2,\ldots,d\}$, considerados primeiro.

Na mistura *a posteriori*

$$p(\theta/r) = \sum_{i=1}^{N} W_i \, C\{\alpha;\, Q_i(r)\} \, h\{Q_i(\theta)\} \prod_{i=1}^{n} \theta_i^{r_i} , \qquad \theta \in S^{(n)}$$

os pesos W_i dados por

$$W_i = \frac{1/C\{\alpha;\, Q_i(r)\}}{\sum_{j=1}^{N} 1/C\{\alpha;\, Q_j(r)\}} \qquad i=1,\ldots,N \qquad (12)$$

indicam o grau com que os dados sustentam a suposição a priori de que o (i_{j+1})-ésimo teste é menos útil do que o (i_j)-ésimo teste (j=1,...,k).

Considerando (12) para α=1 e i=1, temos

$$W_1 = \frac{1}{\left(\sum_{i=1}^{n/2} r_i + 1\right)\left(\sum_{i=(n/2)+1}^{n} r_i + 1\right)} \cdot \frac{1}{\left(\sum_{i=1}^{n/4} r_i + 1\right)\cdots\left(\sum_{i=(3n/4)+1}^{n} r_i + 1\right)} \cdots$$

$$\cdots \frac{1}{\left(\sum_{i=1}^{v} r_i + 1\right)\cdots\left(\sum_{i=n-v+1}^{n} r_i + 1\right)}$$

$$v = 2^{d-k+1}$$

Como já discutido para d=2, vemos que o produto de cada dois termos sucessivos no denominador da expressão acima é aproximadamente proporcional à variância amostral de X_j dado $X_1,\ldots,X_{j-1}$ (j=1,...,k).

Condicionado em $X_1,\ldots,X_{j-1}$, quanto maior a variância amostral de uma variável X_j, ocorrendo o máximo para $\hat{P}(X_j=1/X_1,\ldots X_{j-1}) \tilde{=} 1/2$, menos útil para classificação ela será. Segue-se então que quanto menor for o peso W_1, menos útil será a ordenação de testes $(X_1,\ldots,X_k)$, com conclusão análoga para qualquer $\underset{\sim}{\alpha} > 0$.

Observe, por exemplo, a interpretação de $C_1(\alpha;\ Q(r))$ em d=3, α=1 e k=2

X_1	X_2	X_3	r_i
1	1	1	r_1
1	1	0	r_2
1	0	1	r_3
1	0	0	r_4
0	1	1	r_5
0	1	0	r_6
0	0	1	r_7
0	0	0	r_8

$$n = 2^d = 8$$
$$v = 2^{d-k+1} = 4$$

Para que W_1 seja grande deve ocorrer o inverso com $C_1(\alpha, r)$, associado à ordenação (X_1, X_2, X_3), de forma que $Q_1(r) = (r_1, \ldots, r_8)$.

Temos

$$C_1(\alpha; r) = \frac{(n+1)!}{\prod_{i=1}^{n} \Gamma(r_i + \frac{1}{2})} \left[\sum_{i=1}^{n/2} r_i + 1 \right] \left[\sum_{i=(n/2)+1}^{n} r_i + 1 \right]$$

onde

$$C_1(\alpha;\ r) \propto \frac{\hat{P}(X_1=1)\ \hat{P}(X_1=0)}{\hat{P}(X_1=1, X_2=1, X_3=1)\hat{P}(X_1=1, X_2=1, X_3=0)\ldots\hat{P}(X_1=0, X_2=0, X_3=0)}$$

$$= \frac{\hat{P}(X_1=1)}{\hat{P}(X_2=1, X_3=1/X_1=1)\hat{P}(X_1=1)\hat{P}(X_1=1, X_2=1, X_3=0)} \cdot$$

$$\cdot \frac{1}{\hat{P}(X_1=1, X_2=0, X_3=1)\hat{P}(X_1=1, X_2=0, X_3=0)} \cdot$$

$$\cdot \frac{\hat{P}(X_1=0)}{\hat{P}(X_2=1,X_3=1/X_1=0)\hat{P}(X_1=0)\hat{P}(X_1=0,X_2=1,X_3=0)} \cdot$$

$$\cdot \frac{1}{\hat{P}(X_1=0,X_2=0,X_3=1)\hat{P}(X_1=0,X_2=0,X_3=0)}$$

$$= \frac{1}{\hat{P}(X_2=1,X_3=1/X_1=1)\hat{P}(X_2=1,X_3=0/X_1=1)\hat{P}(X_2=0,X_3=1/X_1=1)} \cdot$$

$$\cdot \frac{1}{\hat{P}(X_2=0,X_3=0/X_1=1)[\hat{P}(X_1=1)]^3} \cdot$$

$$\cdot \frac{1}{\hat{P}(X_2=1,X_3=1/X_1=0)\hat{P}(X_2=1,X_3=0/X_1=0)\hat{P}(X_2=0,X_3=1/X_1=0)} \cdot$$

$$\cdot \frac{1}{\hat{P}(X_2=0,X_3=0/X_1=0)[\hat{P}(X_1=0)]^3}$$

Vemos na expressão acima que $C_1(\alpha;\ r)$ será mínimo quando $\hat{P}(X_2=x_2,X_3=x_3/X_1) \tilde{=} \frac{1}{4}$, o que tornaria máximo o produto do denominador.

Assim, condicionado na observação de X_1 as variáveis X_2 e X_3 seriam pouco valiosas para auxiliar a classificação, como se X_1 fosse "quase suficiente" para cumprir este objetivo.

Poderíamos raciocinar também que o par de testes (X_1,X_2) é "inteiramente informativo", tornando X_3 desprezível, se enxergarmos a expressão $C_1(\alpha;\ r)$ em função de $\hat{P}(X_3=x_3/X_1,X_2)$ e estas probabilidades forem muito próximas de $\frac{1}{2}$.

Assim, entre os $Z=\binom{d}{k}$ sub-grupos de k variáveis, a utilidade de um deles para classificação pode ser expressa pelo produtório

$$\prod_{i\in z} W_i$$

onde Z é o conjunto de k! elementos do conjunto $\{1,\ldots,N=\frac{d!}{(d-k)!}\}$, correspondendo às k! permutações das variáveis X_{Z_j} $(j=1,\ldots,k)$.

Voltando ao problema da classificação do paciente em duas classes D_1 e D_2, consideremos P_1 e P_2 suas probabilidades *a priori*, respectivamente, $P_1 + P_2 = 1$, isto é, $P_1 = P(D_1)$ é a probabilidade *a priori* atribuída a Π_1.

Três grupos de pesos, como definidos acima, $\{W_i^{(1)}\}$, $\{W_i^{(2)}\}$ e $\{W_i^{(0)}\}$, $(i=1,\ldots,N)$ são derivados, sendo os dois primeiros nos dados amostrais de D_1 e D_2, respectivamente, e o último do conjunto inteiro de dados, como se tivessem sido originados da mesma classe.

Em concordância às interpretações vistas acima, pode-se concluir que quanto maior for o peso $W_i^{(0)}$ menos útil seria a ordenação $(X_{i_1},X_{i_2},\ldots,X_{i_n})$ para classificação. Isto ocorre porque se $W_i^{(0)}$ é alto nos dados dos dois grupos reunidos, os sintomas $(X_{i_1},\ldots,X_{i_k})$ teriam sua ocorrência com comportamento muito similar dentro de cada grupo D_1 e D_2. Logo, esta ordenação de testes não daria um subsídio valioso para separar os dois grupos.

Kokolakis (1985) estabelece o seguinte critério para selecionar as melhores k variáveis $(1 \leq k < d)$ entre as d disponíveis.

Para um sub-grupo de k variáveis $\{X_{s_1},\ldots,X_{s_k}\}$ do grupo $\{X_1,\ldots,X_d\}$ considera-se

$$R = \prod_{i \in Z} \frac{\{W_i^{(1)}\}^{P_1} \{W_i^{(2)}\}^{P_2}}{W_i^{(0)}} \qquad (13)$$

onde Z é o grupo de k! elementos do conjunto $\{1,\ldots,N\}$ correspondendo às k! permutações do sub-grupo $\{X_{Z_1},\ldots,X_{Z_n}\}$.

Fixando este números de k variáveis a serem escolhidas, o critério estabelece então a preferência pelo sub-grupo que apresenta o maior valor para *R*.

A título de ilustração observemos o seguinte exemplo.

Exemplo:

d = 3 testes disponíveis

k = 2 testes a serem selecionados

Temos $N = \frac{\Gamma(d+1)}{\Gamma(d-k+1)} = 6$ permutações dos testes X_1, X_2 e X_3.

Evidentemente cada uma destas permutações tem um sentido próprio, quanto à viabilidade da aplicação sucessiva dos testes clínicos ou mesmo em relação à observação dos sintomas.

A utilidade dos testes X_1 e X_2 para classificação dos pacientes seria medida pelo R_{12}, associado aos pesos das permutações (X_1,X_2,X_3) e (X_2,X_1,X_3) nos dois grupos D_1 e D_2 e no global D_0, onde $D_0=D_1 \cup D_2$.

Consideremos hipoteticamente

(a)

$W_1^{(1)} = 0,6$ $\quad W_1^{(2)} = 0,5$ $\quad W_1^{(0)} = 0,53$

$W_2^{(1)} = 0,2$ $\quad W_2^{(2)} = 0,25$ $\quad W_2^{(0)} = 0,23$

Os pesos $W_1^{(1)}$, $W_1^{(2)}$ e $W_1^{(0)}$ seriam associados à permutação (X_1,X_2,X_3) nos grupos D_1, D_2 e D_0, respectivamente, o mesmo ocorrendo para $W_2^{(1)}$, $W_2^{(2)}$ e $W_2^{(0)}$ quanto à permutação (X_2,X_1,X_3).

Assim

(a_1)

P_1	R_{12}
0,5	1,0047
0,4	1,0088
0,1	1,0213

Vemos que embora os indicadores clínicos tenham grande associação entre sí nos grupos D_1 e D_2, este comportamento é similar de forma que os testes clínicos X_1 e X_2 não são úteis para classificação dos pacientes.

Entretanto, mudando os pesos de $W_i^{(0)}$, (i=1,2), para $W_1^{(0)}$= 0,21 e $W_2^{(0)}$= 0,18, temos

(a_2)

P_1	R_{12}
0,5	3,2401
0,4	3,2533
0,1	3,2834

Agora percebe-se a enorme utilidade de X_1 e X_2 para classificação, pois suas distribuições nos grupos D_1 e D_2 estão fortemente associadas mas em sentido inverso, de forma que na distribuição global (D_0) esta associação quase desaparece.

Considerando por outro lado

(b)

$$W_1^{(1)} = 0,40 \qquad W_1^{(2)} = 0,36 \qquad W_1^{(0)} = 0,37$$

$$W_2^{(1)} = 0,28 \qquad W_2^{(2)} = 0,27 \qquad W_2^{(0)} = 0,273$$

temos agora pesos mais próximos do "peso neutro 1/6" (quando $W_i = \frac{1}{N} = \frac{1}{6}$) nos grupos D_1 e D_2. De maneira análoga a (a_1) encontramos

(b_1)

P_1	R_{12}
0.5	1.0329
0.4	1.0184
0.1	0.9760

Redefinindo os pesos de D_0 para $w_1^{(0)}=0.19$ e $w_2^{(0)}=0.16$, temos

(b_2)

P_1	R_{12}
0.5	3.4322
0.4	3.3839
0.1	3.2430

Vemos portanto que R é sensível em relação à diferença entre os pesos de (D_1 e D_2) e D_0. A comparação de (a_2) e (b_2) mostra que embora as sequências (X_1,X_2,X_3) e (X_2,X_1,X_3) de indicadores clínicos tenham menor importância no último caso, as variâveis X_1 e X_2 continuam mantendo uma grande utilidade para classificação, já que o comportamento destes sintomas difere nos grupos D_1 e D_2.

Ao observarmos as tabelas acima não fica claro a forma de influência das prioris P_1 e P_2 nos dados hipotéticos. Certamente uma utilidade crescente das variáveis X_1 e X_2, como em (a), a medida que a classe de doentes D_1 se torna rara, deve estar correspondido a níveis de sensibilidade e especificidade dos dois testes.

Nas simulações feitas adiante poderão ser tiradas algumas conclusões a respeito.

3.4 Comparação com outros critérios de seleção

Embora muitos critérios se prestem para seleção de variáveis binárias, restringiremos a comparação do desempenho do coeficiente de Kokolakis em relação a apenas quatro, definidos abaixo.

3.4.1 Critério de Informação de Akaike (AIC)

A introdução do AIC é baseada no princípio de maximização de entropia: formula-se como objetivo de inferência estatística a estimação da verdadeira distribuição dos dados e tenta-se encontrar a estimativa que maximiza a entropia esperada. Assim, a média do log esperado da verossimilhança seria considerada como uma medida de ajuste de um modelo, sendo que quanto maior seu valor melhor seria o ajuste do modelo.

Entretanto, o máximo do log da verossimilhança tem a tendência geral de sub-estimar o verdadeiro valor da média do log esperado da verossimilhança, sendo esta tendência mais comum em modelos com um grande número de parâmetros.

Logo, o valor de

$$\text{AIC} = -2 \times (\text{máximo do log da verossimilhança do modelo}) + 2 \times (\text{nº de parâmetros livres do modelo})$$

é o critério para seleção de modelo.

Especificamente, no caso de seleção de variáveis para classificação onde o interesse é diagnosticar o indivíduo no grupo D_1 (indicando este fato por h=1) ou D_2 (quando h=0), a partir das variáveis explanatórias binárias $X_1,\ldots,X_d$, teríamos o seguinte. Denotando qualquer sub-grupo do conjunto de indicadores $(X_1,\ldots,X_d)$, o AIC correspondente ao modelo em que a classificação do paciente dependeria apenas de k (k=1,...,d) variáveis segundo Sakamoto *et. al.* (1983) é dado por:

$$AIC\ (H;\ X_1,\ldots,X_k) =$$

$$= (-2) \sum_{h,\ x_1,\ldots,x_k} m(h,\ x_1,\ldots,x_k)\ \log \frac{m.m(h,\ x_1,\ldots,x_k)}{m(h)m(x_1,\ldots,x_k)} +$$

$$+ 2(C_x - 1)\ , \qquad (14)$$

$$h,\ x_1,\ldots,x_k = 0,1$$

onde

$m(x_1,\ldots,x_n)$ é frequência marginal de uma combinação das variáveis explanatórias

C_x: n^{o} de categorias relativas ao sub-grupo $(X_1,\ldots,X_k)$, logo, $C_x = 2^k$ assumindo-se $m(\phi) = m$

$$C_\phi = 1$$

Assim, para um número fixado de k variáveis, seria escolhido aquele sub-grupo com menor AIC, indicando então os melhores indicadores para estabelecer um modelo de classificação, a partir do relacionamento entre a posse ou não da doença e os indicadores binários.

3.4.2 Medida de Informação de Lindley e Divergência de Jeffreys

Para uma dada função ϕ definida em um intervalo I, a ϕ-entropia de $\mathbf{p} = (p_1,\ldots,p_n) \in I^n$ é definida (Burbea e Rao(1982)) como

$$H_{n,\phi}(\mathbf{p}) = - \sum_{i=1}^{n} \phi(p_i), \quad \mathbf{p} \in I^n$$

e para dois vetores $\mathbf{p}^{(1)}$, $\mathbf{p}^{(2)} \in I^n$, a diferença de Jensen

$$J_{n,\phi}(\mathbf{p}^{(1)},\mathbf{p}^{(2)}) = H_{n,\phi}\left[\frac{\mathbf{p}^{(1)}+\mathbf{p}^{(2)}}{2}\right] - \frac{1}{2}\left[H_{n,\phi}(\mathbf{p}^{(1)}) + H_{n,\phi}(\mathbf{p}^{(2)})\right]$$

baseada na ϕ-entropia torna-se

$$J_{n,\phi}(\mathbf{p}^{(1)},\mathbf{p}^{(2)}) = \sum_{i=1}^{n}\left\{\frac{1}{2}\left[\phi(p_i^{(1)}) + \phi(p_i^{(2)})\right] - \phi\left[\frac{p_i^{(1)}+p_i^{(2)}}{2}\right]\right\}$$

$$(\mathbf{p}^{(1)},\mathbf{p}^{(2)}) \in I^n \times I^n$$

Quando $\phi(x) = x \log x$, a diferença de Jensen reduz-se à *Medida de Informação de Lindley* fornecida por um experimento

$$L = H(\mathbf{p}^{(0)}) - P_1 H(\mathbf{p}^{(1)}) - P_2 H(\mathbf{p}^{(2)}) \qquad (15)$$

Baseado em entropias de grau α, $H_{n,\alpha}(\mathbf{p})$, com

$$\phi_\alpha(\mathbf{p}) = \begin{cases}(\alpha - 1)^{-1}(\mathbf{p}^\alpha - p) & \alpha \neq 1\\ p \log p & \alpha = 1\end{cases}$$

a medida de divergência de Kullback e Leibler para $\alpha = 1$ equivale à *Divergência de Jeffreys*

$$J = \sum_{i=1}^{n}(p_i^{(1)} - p_i^{(2)})\,[\log p_i^{(1)} - \log p_i^{(2)}] \qquad (16)$$

3.4.3 Coeficiente de Bhattacharyya

Sejam duas populações multinomiais caracterizadas por dois vetores de probabilidade $(p_1^{(1)},\ldots,p_n^{(1)})$ e $(p_1^{(2)},\ldots,p_n^{(2)})$. Então, como $\sum_{i=1}^{n} p_i^{(1)} = 1$ e $\sum_{i=1}^{n} p_i^{(2)} = 1$, $\left(\sqrt{p_1^{(1)}},\ldots,\sqrt{p_n^{(1)}}\right)$ e $\left(\sqrt{p_1^{(2)}},\ldots,\sqrt{p_n^{(2)}}\right)$ podem ser considerados os cossenos diretores das duas retas

passando pela origem em um espaço n-dimensional.

O quadrado do ângulo entre estas duas linhas pode ser considerado uma medida apropriada de divergência entre estas duas populações(Bhattacharyya(1946)).

Assim, se a medida de divergência é denotada por Δ^2, então

$$\cos \Delta = \sqrt{p_1^{(1)} p_1^{(2)}} + \ldots + \sqrt{p_n^{(1)} p_n^{(2)}} .$$

A partir desta medida, Matusita(1951) apresenta o coeficiente $B = - \log \rho$, onde

$$\rho = \left[\sum_{i=1}^{n} p_i^{(1)} p_i^{(2)} \right]^{1/2} \qquad (17)$$

como uma medida de afinidade entre as duas distribuições.

Para as últimas três medidas, Lindley, Jeffreys e Bhattacharyya, um grande valor para o coeficiente indica que a distribuição conjunta dos indicadores clínicos em cada grupo de pacientes (portadores e não-portadores da doença) diverge bastante. Logo, a observação destas variáveis em um paciente se constitui num subsídio valioso para diagnóstico.

4. COMPARAÇÕES POR SIMULAÇÕES EM MONTE CARLO

4.1 Introdução

O objetivo das simulações é observar o comportamento do desempenho do coeficiente de Kokolakis na seleção das melhores variáveis para classificação, frente a outros critérios, aplicado a particulares estruturas de dependência entre os testes. Estas simulações foram realizadas através de um programa em linguagem Fortran no VAX da Unicamp, sendo os resultados apresentados no capitulo 5.

Consideremos o caso especial onde seria buscada a seleção de duas variáveis em três disponíveis para classificação dos pacientes. O motivo da redução deste número de variáveis, como já foi discutido anteriormente, seria no sentido de reduzir o custo da aplicação dos testes clínicos nos pacientes ou outro associado à rapidez na coleta dos dados possibilitando também a construção de regras parcimoniosas de diagnóstico.

Suponhamos então a disponibilidade dos testes clínicos X_1, X_2 e X_3, que indicam a presença ou não de três características distintas associadas a uma determinada doença de interesse, com as seguintes distribuições:

X_1	$P(X_1/D_1)$	$P(X_1/D_2)$
1	0.9	0.2
0	0.1	0.8

X_2	$P(X_2/D_1)$	$P(X_2/D_1)$
1	0.55	0.48
0	0.45	0.52

X_3	$P(X_3/D_1)$	$P(X_3/D_2)$
1	0.75	0.3
0	0.25	0.7

Observando estas distribuições acima e as distribuições conjuntas de (X_1, X_3), sob independência nas classes D_1 e D_2

$P(X_1, X_3/D_1)$

X_1 \ X_3	1	0
1	0.675	0.225
0	0.075	0.025

$P(X_1, X_3/D_2)$

X_1 \ X_3	1	0
1	0.06	0.14
0	0.24	0.56

parece evidente que os testes X_1 e X_3 constituem o melhor par de indicadores para classificar um paciente, sendo o teste X_2 praticamente inútil para auxiliar no diagnóstico.

A observação da presença conjunta das caracteristicas X_1 e X_3, isto é, $(x_1, x_3)=(1,1)$, ou a ausência conjunta de ambas, aparenta ser uma evidência valiosa para diagnosticar o paciente como integrante da classe D_1 ou D_2, respectivamente.

Entretanto, ao utilizarmos a taxa de erro como critério de desempenho das regras de classificação, sem o uso de uma função utilidade com conseqüências diferenciadas para o *falso positivo* ou *falso negativo*, a regra de Bayes pode criar um par melhor do que este acima.

Para a situação de independência conjunta entre as três variáveis em ambos os grupos, teriamos as taxas ótimas de erro para variáveis e pares segundo niveis de prevalência da doença, conforme tabela abaixo.

	Π_1				
F	0.1	0.2	0.3	0.4	0.5
X_1	0.10	0.18	0.17	0.16	0.15
X_2	0.10	0.20	0.30	0.40	0.465
X_3	0.10	0.20	0.285	0.28	0.275
X_1, X_2	0.10	0.1778	0.17	0.16	0.15
X_1, X_3	0.0865	0.113	0.1395	0.16	0.15
X_2, X_3	0.10	0.20	0.27705	0.28	0.275

Inicialmente percebemos que sob o critério da taxa de erro de classificação as variáveis X_1 e X_3 se constituem sempre nos melhores testes individuais, na região de incidência (ou prevalência) considerada da doença (isto é, o tamanho da classe D_1).

Entretanto, dependendo da prevalência da doença sob estudo, o par (X_1,X_2) pode competir com (X_1,X_3), apresentando o mesmo desempenho. Nesta situação, dependendo do esforço ou custo para observar as variáveis X_2 e X_3, o pesquisador poderia optar pela utilização de X_1 e X_2. Contudo, este não parece ser ainda o caso onde este deveria ser o par de variáveis a ser escolhido pois X_2 não consegue nunca contribuir de forma significativa na redução da taxa individual de erro sob utilização de X_1.

Vemos também que a medida que a prevalência da doença diminui não se distingue a melhor variável individual ou par de testes para classificação.

Não estudaremos aqui diferentes níveis de sensibilidade e especificidade para que os testes criem situações mais ricas e conclusivas a respeito de doenças raras, para observar o desempenho do coeficiente de Kokolakis.

Assim, serão consideradas nove formas de estruturas de dependência entre os três testes, para suas distribuições unidimensionais fixadas no início. Em relação às taxas de prevalência da doença serão considerados os níveis Π_1=0.5, 0.25 e 0.10.

4.2 As estruturas simuladas de dependência entre os indicadores

As estruturas criadas foram Independência entre os três testes e outras oito situações de dependência entre duas variáveis. Ao utilizarmos a expressão (4) para a taxa de erro ótima dada pelo par (X_1, X_2) vemos que sua taxa não é alterada por qualquer valor da diferença das covariâncias entre estas variáveis nos grupos D_1 e D_2, no caso $\Pi_1=\Pi_2$.

Criamos assim oito situações com diferentes valores para as covariâncias entre X_1 e X_3 ou X_2 e X_3 nas classes, de forma a tornar o par (X_1, X_3) como aquele com menor taxa ótima de erro ou então diminuir a taxa de erro do par (X_2, X_3) para que o mesmo atrapalhe um pouco a superioridade dos outros dois pares de testes.

Entretanto, como estas estruturas de covariância também foram estendidas aos casos onde Π_1=0.25 e 0.10, diferentes situações irão surgir, chegando mesmo a inverter a ordem de superioridade entre os três pares.

Não é nossa preocupação explicar aqui a viabilidade dos valores das covariâncias, quando poderemos assumir, por exemplo, uma forte correlação negativa entre dois sintomas no grupo de indivíduos portadores da doença.

Na classificação de nascimentos de crianças, por exemplo, poderiam ser consideradas "normais" aquelas com pequeno tempo de gestação e baixo peso ao nascer ou com tempo de gestação longo e alto peso ao nascer. Enquanto isso, crianças "não normais" seriam aquelas de longa gestação mas baixo peso ao nascer ou com curta gestação e alto peso ao nascer.

Logo, dependendo da aplicação prática, pode ocorrer que determinadas estruturas de dependência entre as características ou sintomas levem a regra a classificar um indivíduo com indicadores $(x_i, x_j)=(0,0)$ e $(1,1)$ em um grupo, e $(0,1)$ e $(1,0)$ em outro, embora isto não pareça comum.

Temos então as seguintes estruturas estudadas, acompanhadas pelas taxas ótimas de erro individuais e para pares de testes, para os três níveis de prevalência estudados:.

e_1: independência

	Π_1		
F	0.10	0.25	0.5
X_1	0.10	0.175	0.15
X_2	0.10	0.25	0.465
X_3	0.10	0.25	0.275
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.10	0.25	0.275
X_1, X_2, X_3	0.0865	0.12625	0.15

e_2: ($\rho_{13} = 0.05$, $\tau_{13} = 0.05$)

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.10	0.15125	0.15
X_2, X_3	0.10	0.25	0.275
X_1, X_2, X_3	0.10	0.15125	0.15

e_3: ($\rho_{13} = 0.03$, $\tau_{13} = -0.04$)

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.11	0.175	0.15
X_1, X_3	0.0575	0.08875	0.15
X_2, X_3	0.11	0.25	0.275
X_1, X_2, X_3	0.0575	0.08875	0.14708

e_4: ($\rho_{13} = 0.05$, $\tau_{13} = -0.045$)

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.041	0.08	0.145
X_2, X_3	0.10	0.25	0.275
X_1, X_2, X_3	0.041	0.08	0.14128

e_5: $(\rho_{23} = 0.10,\ \tau_{23} = -0.08)$

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.10	0.16988	0.275
X_1, X_2, X_3	0.0654	0.12104	0.12525

e_6: $(\rho_{23} = -0.10,\ \tau_{23} = 0.13)$

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.07965	0.16013	0.25925
X_1, X_2, X_3	0.06531	0.10371	0.09475

e_7: $(\rho_{23} = 0.12,\ \tau_{23} = -0.11)$

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.07735	0.14238	0.25075
X_1, X_2, X_3	0.0582	0.11204	0.10025

e_8: (ρ_{23} = -0.08, τ_{23} = 0.14)

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.07065	0.15263	0.25425
X_1, X_2, X_3	0.06351	0.10088	0.08975

e_9: (ρ_{23} = 0.10, τ_{23} = -0.13)

	Π_1		
F	0.1	0.25	0.5
X_1, X_2	0.10	0.175	0.15
X_1, X_3	0.0865	0.12625	0.15
X_2, X_3	0.06135	0.13238	0.25075
X_1, X_2, X_3	0.0564	0.10913	0.10025

4.3 Dos tamanhos amostrais e número de simulações

Para cada situação, isto é, uma estrutura associada com um nível de prevalência da doença, foram tomadas 500 amostras simuladas de pacientes. O processo foi amostragem aleatória simples na população $D_0 = D_1 \cup D_2$ de interesse, de forma que o número esperado de pacientes com a doença na amostra era $m\Pi_1$, onde m era o tamanho amostral fixado.

Em cada situação foram tomados três tamanhos diferentes de amostra, fixados em 100, 250 e 360 pacientes, para acompanhar o desempenho dos coeficientes sob comparação de forma sucessiva. O último tamanho amostral, 360, foi tomado para satisfazer o requisito citado por Stanish e Allred(1981). Em uma tabela multidimensional, onde cada sub-tabela representa a classificação cruzada da variável dependente (grupo ou classe do paciente, D_1 ou D_2) com a covariável que está sendo considerada para inclusão no modelo, esta pode ser vista como um conjunto de sub-tabelas rxc.

O teste da razão de verossimilhança(TRV) para inclusão de variáveis que causem a maior redução na entropia somente seria apropriado quando o tamanho amostral de cada sub-tabela fosse suficientemente grande (m maior que 20 para cada grau de liberdade extraído dos dados). Ao aplicarmos este conceito às estruturas acima, uma delas exigiu 363 pacientes, sendo este número o máximo solicitado. Assim, foi este o máximo assumido para todas, onde então haveria estabilidade para a inclusão de variáveis no modelo sob o TRV.

Teremos então para cada uma das 27 situações (9 estruturas de dependência das variáveis combinadas com 3 níveis de prevalência da doença na população) 500 amostras aleatórias simples de tamanhos 100, 250 e 360 pacientes.

4.4 Sobre o critério para desempenho dos coeficientes

Em cada uma das situações, para cada tamanho amostral será registrado o número de vezes nas 500 amostras geradas em que os coeficientes de Kokolakis, Lindley, Jeffreys, Akaike e Bhattacharyya são máximos para o par (X_1, X_3), indicando (ou detectando) que estes seriam os melhores indicadores para classificação.

A opção por este par deve-se ao fato de que em todas as estruturas simuladas sob prevalências 0.50 e 0.25, o mesmo é composto pelas melhores variáveis individuais e sua taxa de erro comum é não superior aos outros dois pares. Além disso, sem considerarmos o custo

de observação das variáveis, no caso de valor perdido ou informação não registrada de um destes dois indicadores num paciente, o outro restante permitiria um diagnóstico ainda satisfatório, com taxa de erro não significativamente aumentada.

Assim, na estrutura e_3 com prevalência 0.25, por exemplo, para confirmar sua utilidade de seleção de variáveis, todos os coeficientes investigados deverão detectar o par (X_1, X_3) como o melhor em quase todas as 500 amostras geradas para cada um dos três tamanhos amostrais pois sua taxa ótima de erro é minima.

Contudo, sob prevalência 0.10, este par não se constitui no melhor (com menor taxa ótima de erro) nas estruturas e_6, e_7, e_8 e e_9. Neste caso, embora as simulações possam criar particulares amostras onde (X_1, X_3) se mostre o melhor par, espera-se que um bom coeficiente selecione este par um número muito pequeno de vezes entre as 500 simulações. Obviamente as taxas de erro dos pares (X_1, X_3) e (X_2, X_3) são relativamente próximas, mas os coeficientes deverão ser sensíveis para a superioridade deste último par, que apresenta a menor taxa ótima de erro.

Para os coeficientes de Lindley e de Kokolakis veremos também o efeito da probabilidade a priori atribuída aos dois grupos, supondo que o pesquisador desconheça a prevalência que a doença exerce na população sob estudo.

Logo, para cada nível de prevalência em que as simulações serão feitas consideraremos o poder de detecção do par (X_1, X_3) pelos coeficientes citados, utilizando as prioris de 0.10, 0.25 e 0.50. Veremos então o desempenho de cada coeficiente ao especificar uma probabilidade a priori P_1 que concorda com a prevalência Π_1 da doença ou que comete uma alta sub-estimação ou super-estimação de seu valor.

Os coeficientes de Bhattacharyya, Jeffreys e Lindley foram calculados através de distribuições preditivas baseadas em distribuições a priori dadas como em (11).

4.5 Os modelos de classificação e seu desempenho

As três regras de classificação que usaremos aqui para um paciente com um vetor de indicadores observados $\mathbf{x}=(x_1,\ldots,x_d)$ proveniente da classe D_i $(i=1,2)$ com probabilidade a priori P_i serão:

Modelo 1: Kokolakis

Aloca-se $\mathbf{x}$ a D_1 se

$$P_1 \frac{R_x^{(1)} + \alpha_d}{m_1 + 2\alpha_d} \sum_{i=1}^{d!} V_i^{(1)}(\mathbf{x}) W_i^{(1)} \geq P_2 \frac{R_x^{(2)} + \alpha_d}{m_2 + 2\alpha_d} \sum_{i=1}^{d!} V_i^{(2)}(\mathbf{x}) W_i^{(2)} \qquad (18)$$

e, caso contrário, a D_2.

Segundo Kokolakis(1983) prioris do tipo (7) com g(Θ) definida por (8) não são coerentes com cada uma das outras para diferentes dimensionalidades do vetor de indicadores. Entretanto, pode-se atingir uma coerência aproximada requerendo-se que a priori as variâncias de $P(X_i=1)$ $(i=1,\ldots,d)$ para d fixado sejam todas iguais a

$$V(\alpha,d) = V[P(X_i=1)] = \frac{1}{4\alpha d}\left[1 - \left(\frac{1+\alpha}{1+2\alpha}\right)^d\right] \qquad (i=1,\ldots,d)$$

podendo permanecer a mesma se indicadores são introduzidos ou deletados.

Como $V(\alpha,d)$ é decrescente tanto em α como em d, um decréscimo apropriado de α precisa acompanhar um acréscimo de d devido à introdução de indicadores adicionais.

Modelo 2: Dirichlet Simétrica

Aloca-se $\mathbf{x}$ a D_1 se

$$P_1 \frac{R_X^{(1)} + 2/n}{m_1 + 2} \geq P_2 \frac{R_X^{(2)} + 2/n}{m_2 + 2} \qquad (19)$$

e, caso contrário, a D_2.

Modelo 3: Independência

Embora possa ser inadequado ou irrealista, foi encontrado por Titterington *et. al.*(1981) como um dos melhores modelos para classificação.

Assim, aloca-se X a D_1 se

$$P_1 \prod_{i=1}^{d} \frac{R_{X_i}^{(1)} + 1/2}{R_{X_i}^{(1)} + R_{\bar{X}_i}^{(1)} + 1} \geq P_2 \prod_{i=1}^{d} \frac{R_{X_i}^{(2)} + 1/2}{R_{X_i}^{(2)} + R_{\bar{X}_i}^{(2)} + 1} \qquad (20)$$

e, caso contrário, a D_2, com $\bar{x}_i = 1 - x_i$ $(i=1,\ldots,d)$ e $R_{X_1}^{(j)},\ldots,_{X_l}$ $(j=1,2\ ;\ l=1,\ldots,d)$ a frequência do vetor $(x_1,\ldots,x_l)$ nos dados da amostra piloto da classe D_j.

Este modelo de classificação é equivalente a assumir que as chances $P(X_i=1)$ $(i=1,\ldots,d)$ são independentes com probabilidades a priori simétricas *beta* de parâmetro $\alpha = \frac{1}{2}$.

Por motivos de esforço computacional as três regras de classificação acima foram estabelecidas somente para o caso em que Π_1 era conhecida.

Para cada uma das situações (estrutura associada ao nível de prevalência da doença), sob tamanho amostral fixado, os três modelos de classificação foram aplicados aos pacientes em cada amostra das 500 simuladas.

A comparação entre eles deu-se pelas taxas atuais de erro, registrando-se sob cada par de variáveis empregadas para classificação o número de amostras em que cada modelo obtém *falso positivo* e *falso negativo* não superior aos demais pares.

Assim, por exemplo, um modelo que registrasse, numa particular situação, 487 vezes um falso positivo e 432 vezes um falso negativo não superiores aos correspondentes falso positivo e falso negativo dados pelos outros dois modelos, nas 500 amostras simuladas, estaria apresentando um ótimo desempenho ao utilizar aquele par de variáveis para emitir diagnóstico.

Além disso, foi registrado o desempenho de cada modelo na *amostra futura*, ou seja, a regra de classificação criada na t-ésima amostra foi aplicada na (t+1)-ésima amostra. Para a regra construída na última amostra (t=500) foi gerada uma amostra adicional com o objetivo exclusivo de aplicação desta regra existente, para cada modelo de classificação.

Tivemos assim a comparação de cada modelo através da taxa de falso positivo e falso negativo na amostra atual e na amostra futura, para cada par de indicadores utilizados para classificação. Ao final, além do registro do número de amostras em que cada modelo se mostrou não inferior aos demais, foi observado também a *taxa média de falso positivo e falso negativo*, para cada modelo, sob cada par de variáveis empregado.

Devido à baixa dimensionalidade do vetor de informações para diagnóstico, as observações $(x_i, x_j)=(1,1)$, $(1,0)$, $(0,1)$ ou $(0,0)$ poderiam tornar os modelos pouco distintos entre si, em relação à regra de classificação para cada evidência observada no paciente.

Assim, registrou-se também em quantas vezes das 2000 possíveis (500 amostras com 4 evidências possíveis) o modelo 1 de Kokolakis concordava com o modelo 2 e com o modelo 3 na regra de

classificação, em todas as situações simuladas. Associado a isto, anotou-se também o número de pacientes que foram diagnosticados similarmente, em correspondência à igualdade dos modelos quanto à regra estabelecida para cada par utilizado de indicadores.

5. APRESENTAÇÃO E DISCUSSÃO DOS RESULTADOS

As tabelas 1 a 9 que se seguem apresentam os resultados quanto ao número de vezes em 500 amostras geradas em que os coeficientes estudados detectaram o par (X_1, X_3) como o melhor para classificação.

Inicialmente, observando a tabela 1, vemos que sob independência os coeficientes apresentaram um bom desempenho para todos os níveis de prevalência da doença.

Tabela 1: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_1, por nível de prevalência da doença, tamanho amostral e priori.

	π_1=0.10			π_1=0.25			π_1=0.50		
	m			m			m		
Coeficiente	100	250	360	100	250	360	100	250	360
Akaike	447	491	500	476	500	500	469	500	500
Bhattacharyya	436	481	492	467	498	499	467	498	499
Jeffreys	457	463	474	469	487	498	477	500	500
0.10	403	482	498	442	495	499	347	428	457
L 0.25	320	427	459	461	499	499	407	481	488
0.50	256	362	409	416	484	498	462	500	500
0.10	456	496	500	499	500	500	481	499	500
R 0.25	397	466	481	498	500	500	492	500	500
0.50	305	403	435	477	495	500	499	500	500

(Nota: L indica o coeficiente de Lindley, L 0.10 este mesmo coeficiente sob priori 0.10, R o coeficiente de Kokolakis, e assim por diante)

Vemos que, sob independência e prevalência Π_1=0.10, os coeficientes de Lindley e de Kokolakis apresentam um desempenho

inferior quando utilizam uma priori que super-estima a prevalência da doença, embora o par fixado (X_1, X_3) seja aquele com menor taxa ótima de erro.

Para as outras duas prevalências percebe-se que, embora os demais coeficientes apresentem uma boa detecção do par citado, o desempenho do coeficiente *R* é ligeiramente melhor aos demais, independente da priori utilizada.

Quando colocamos covariâncias entre X_1 e X_3 iguais nas duas classes D_1 e D_2, não alterando as taxas ótimas de erro sob prevalência 0.50 e sem formar um par melhor do que (X_1, X_3) para os outros dois níveis de Π_1, percebemos uma alteração do desempenho, como mostra a tabela 2.

Tabela 2: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_2, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		325	407	426	350	438	467	359	429	453
Bhattacharyya		137	79	58	112	56	30	117	56	32
Jeffreys		344	322	332	323	375	400	378	422	450
	0.10	159	286	358	169	215	246	130	97	75
L	0.25	205	306	356	229	363	415	146	133	121
	0.50	210	314	358	299	411	440	253	365	404
	0.10	465	498	500	498	500	500	482	500	500
R	0.25	403	468	483	491	500	500	495	500	500
	0.50	324	410	445	471	497	500	498	500	500

Os resultados acima mostram a grande vantagem de utilizar o coeficiente *R* para selecionar o melhor par de variáveis. Mesmo que

outros pares apresentem mesma taxa ótima de erro sob prevalências 0.10 e 0.50, insistimos na definição de (X_1, X_3) como o melhor par de indicadores, como já reiterado antes.

Entretanto, ao colocarmos covariâncias baixas com sentidos opostos nos grupos, embora o par (X_1, X_3) seja realmente o melhor, volta a ocorrer um desempenho próximo ao da situação de independência.

Tabela 3: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_3, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		490	500	500	496	500	500	492	500	500
Bhattacharyya		431	486	496	484	499	500	492	500	500
Jeffreys		464	475	484	487	497	500	494	500	500
	0.10	474	500	500	496	500	500	407	472	492
L	0.25	403	487	499	496	500	500	449	492	498
	0.50	284	409	425	482	500	500	489	500	500
	0.10	459	495	500	499	500	500	482	500	500
R	0.25	366	453	480	496	499	500	493	500	500
	0.50	268	356	371	471	498	499	499	500	500

A tabela 3 mostra que, como ocorreu para a estrutura e_1, sob prevalência 0.10, o coeficiente de Kokolakis reduz seu desempenho ao super-estimar a prevalência da doença. Contudo, seu desempenho é excelente sob as outras prevalências, sendo equivalente aos demais.

Resultados praticamente análogos aos da última estrutura são obtidos para covariâncias ainda baixas, quase iguais às fixadas anteriormente, como mostra a tabela 4 abaixo.

Tabela 4: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_4, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		498	500	500	499	500	500	497	500	500
Bhattacharyya		460	495	499	486	499	500	489	500	500
Jeffreys		484	493	496	494	499	500	497	500	500
	0.10	492	500	500	495	500	500	410	473	489
L	0.25	447	498	500	498	500	500	452	494	499
	0.50	374	470	495	493	500	500	495	500	500
	0.10	472	499	500	499	500	500	483	500	500
R	0.25	396	474	486	493	499	500	493	500	500
	0.50	305	425	439	470	497	500	500	500	500

Covariâncias altas e em sentido oposto nas classes, como simulado na estrutura e_5, mostram a superioridade do coeficiente de Kokolakis. Como podemos ver adiante na tabela 5, o par (X_1, X_3), que apresenta menor taxa de erro sob as prevalências fixadas é melhor detectado pelo *R*, principalmente sob pequeno tamanho amostral.

Tabela 5: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_5, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		410	479	494	453	499	500	462	500	500
Bhattacharyya		410	471	492	457	498	500	460	496	498
Jeffreys		437	463	469	459	487	496	464	499	500
	0.10	367	478	494	424	491	497	302	398	447
L	0.25	289	411	460	444	499	500	369	469	486
	0.50	239	324	394	408	486	500	462	499	500
	0.10	475	498	500	499	500	500	466	497	500
R	0.25	400	476	489	498	500	500	487	500	500
	0.50	326	419	459	485	500	500	500	500	500

Mantendo ainda altas as covariâncias mas invertendo seu sinal nos grupos em relação à última estrutura, a superioridade do *R* aumenta significativamente sobre os demais coeficientes.

Entretanto, sob prevalência 0.10 o melhor par de indicadores já não é mais (X_1, X_3) mas sim (X_2, X_3), exatamente aquele composto pelas piores variáveis individuais para classificação. Assim, mesmo que as taxas ótimas de erro destes pares sejam próximas e os demais coeficientes cometam o mesmo equívoco quanto à seleção de (X_1, X_3), surpreende o fato de Kokolakis apresentar uma detecção tão alta, como mostra a tabela 6.

Tabela 6: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_6, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		345	414	434	387	457	481	408	479	487
Bhattacharyya		383	465	481	432	488	499	442	496	499
Jeffreys		383	377	365	391	369	381	392	411	400
	0.10	309	425	444	325	387	393	185	197	190
L	0.25	280	401	438	394	470	487	263	316	331
	0.50	264	357	391	391	470	492	423	485	493
	0.10	451	496	499	498	500	500	484	500	500
R	0.25	394	460	479	490	498	500	497	500	500
	0.50	321	408	444	465	496	500	497	500	500

Trabalhando ainda com covariâncias altas nos grupos, colocadas de forma a tornar menor o erro do par (X_2, X_3) sob prevalências baixas, chega-se a resultados similares aos encontrados para a estrutura e_6.

Observando as tabelas que se seguem, quando o par (X_2, X_3) passa a ser o melhor sob π_1=0.10, o coeficiente *R* continua acusando (X_1, X_3) como o par ideal de variáveis num número excessivo de amostras.

Tabela 7: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_7, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		312	390	416	380	455	461	403	469	479
Bhattacharyya		368	462	476	420	483	493	425	481	494
Jeffreys		405	434	420	407	417	419	399	417	440
	0.10	274	398	423	330	409	427	242	288	337
L	0.25	244	373	409	381	464	472	308	378	428
	0.50	220	319	370	368	466	481	409	475	485
	0.10	477	499	500	496	500	500	468	497	500
R	0.25	401	479	493	499	500	500	486	499	500
	0.50	316	426	458	487	500	500	500	500	500

Tabela 8: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_8, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		292	344	390	363	425	441	373	447	464
Brattacharyya		383	449	479	434	486	497	426	493	498
Jeffreys		382	334	349	361	334	311	362	346	336
	0.10	288	354	404	284	290	286	130	111	101
L	0.25	270	373	441	383	440	464	214	237	238
	0.50	258	348	382	388	471	492	412	463	470
	0.10	459	492	497	498	500	500	485	500	500
R	0.25	388	452	474	493	500	500	499	500	500
	0.50	310	394	428	469	497	499	499	500	500

Tabela 9: Detecção do par (X_1, X_3) pelos coeficientes sob estrutura e_9, por nível de prevalência da doença, tamanho amostral e priori.

		π_1=0.10			π_1=0.25			π_1=0.50		
		m			m			m		
Coeficiente		100	250	360	100	250	360	100	250	360
Akaike		268	322	324	339	407	441	401	465	487
Brattacharyya		369	448	462	417	482	492	426	478	491
Jeffreys		383	423	411	395	431	448	389	415	417
	0.10	246	317	334	267	286	276	167	171	139
L	0.25	259	376	409	351	420	446	245	312	315
	0.50	237	346	391	374	375	488	422	469	487
	0.10	469	500	500	497	500	500	473	496	499
R	0.25	411	482	491	494	500	500	486	500	500
	0.50	309	428	462	482	500	500	497	500	500

Todos os coeficientes estudados expressam a utilidade do par de variáveis (X_1, X_3) para classificação, ou seja, quanto maior for o valor assumido por um coeficiente maior será a importância dos indicadores para emitir diagnóstico, com exceção do critério de Akaike, expresso em unidade negativa.

Assim, seria esperada uma correlação expressiva entre eles, já que são guiados pelo mesmo princípio.

No caso de independência sob prevalência 0.50, os resultados não mostram tal fato, mesmo sendo (X_1, X_3) o melhor par de testes, de forma que todos os indicadores deveriam simultaneamente apresentar altos valores para cada amostra simulada.

As tabelas abaixo mostram as correlações estimadas para a situação citada, sob os três tamanhos amostrais fixados.

Sob m_1=100

	B	J	$L_{0.10}$	$L_{0.25}$	$L_{0.50}$	$R_{0.10}$	$R_{0.25}$	$R_{0.50}$
A	-0.949	-0.990	-0.658	-0.807	-0.975	-0.509	-0.494	-0.372
B		0.941	0.761	0.874	0.944	0.630	0.620	0.481
J			0.657	0.801	0.960	0.523	0.503	0.369
$L_{0.10}$				0.973	0.687	0.803	0.724	0.434
$L_{0.25}$					0.836	0.781	0.719	0.462
$L_{0.50}$						0.552	0.544	0.424
$R_{0.10}$							0.960	0.690
$R_{0.25}$								0.866

Sob m_2=250

	B	J	$L_{0.10}$	$L_{0.25}$	$L_{0.50}$	$R_{0.10}$	$R_{0.25}$	$R_{0.50}$
A	-0.963	-0.971	-0.702	-0.839	-0.980	-0.637	-0.661	-0.604
B		0.928	0.800	0.900	0.948	0.722	0.741	0.659
J			0.699	0.802	0.944	0.613	0.642	0.597
$L_{0.10}$				0.974	0.723	0.880	0.818	0.563
$L_{0.25}$					0.860	0.865	0.827	0.620
$L_{0.50}$						0.659	0.685	0.627
$R_{0.10}$							0.974	0.767
$R_{0.25}$								0.892

Sob m_3=360

	B	J	$L_{0.10}$	$L_{0.25}$	$L_{0.50}$	$R_{0.10}$	$R_{0.25}$	$R_{0.50}$
A	-0.966	-0.969	-0.715	-0.846	-0.983	-0.650	-0.664	-0.599
B		0.930	0.814	0.910	0.952	0.748	0.757	0.668
J			0.684	0.813	0.951	0.637	0.655	0.600
$L_{0.10}$				0.975	0.733	0.888	0.826	0.583
$L_{0.25}$					0.866	0.871	0.830	0.630
$L_{0.50}$						0.668	0.683	0.615
$R_{0.10}$							0.978	0.793
$R_{0.25}$								0.904

Vemos acima a evolução crescente das correlações, na medida em que o tamanho amostral aumenta. Na verdade estes resultaddos apenas exibem a grande concordância dos coeficientes em identificar os indicadores (X_1, X_3) para classificação, como visto na tabela 1.

Entretanto, cabe ressaltar a baixa correlação que o coeficiente *R*, sob priori 0.50, apresenta com os coeficientes de Akaike, Bhattacharyya, Jeffreys e Lindley (este, sob a mesma priori 0.50). Embora todos eles tenham uma ótima detecção do melhor par, na medida que uma amostra muito informativa reforçasse a superioridade de (X_1, X_3) para classificação, todos estes coeficientes deveriam assumir um alto valor. Além disso, segundo Kokolakis(1985), seu critério e o de Lindley comportam-se similarmente, com exceção de vetores extremos de probabilidades preditivas.

Ainda sob a estrutura de independência com prevalência 0.50, o coeficiente *R* apresentou uma variabilidade muito baixa, para todos os três níveis de priori fixados.

Para a priori fixada 0.50, combinando portanto com a verdadeira prevalência da doença, vemos no histograma e informações abaixo o comportamento estável deste coeficiente, para o tamanho amostral m=360 a partir da utilização de (X_1, X_3).

Momentos Amostrais do R_{13}

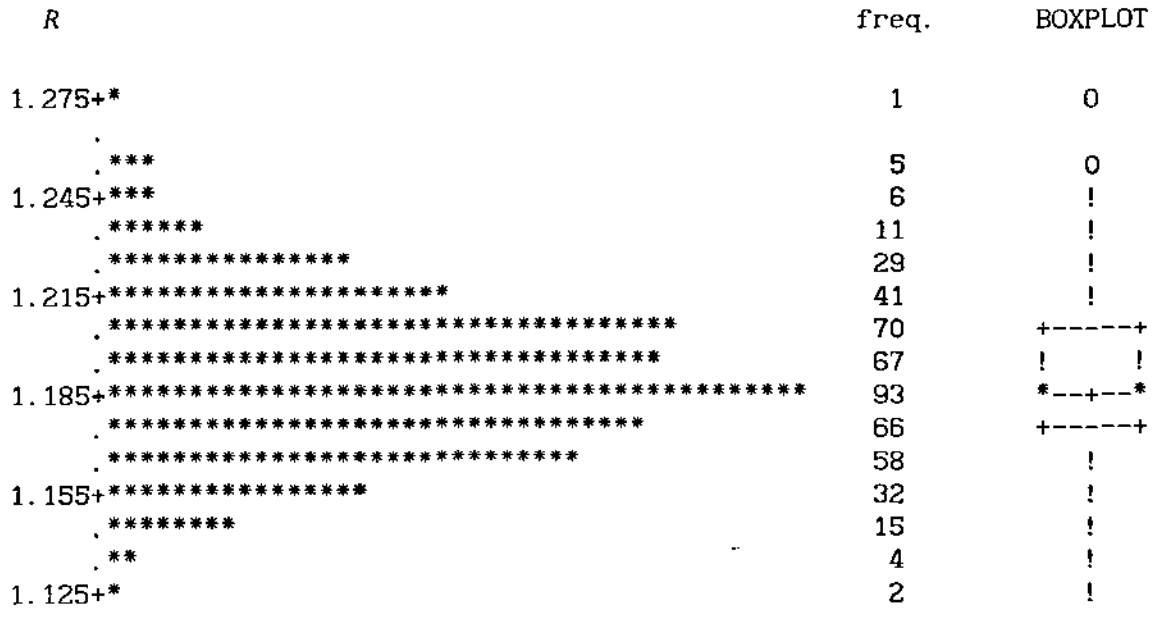
Média: 1.18936 Assimetria: 0.247248

Desvio padrão: 0.0242004 Kurtose: 0.0341883

Coeficiente de Variação: 2.03473

Quantis

100% max	1.27146	99%	1.25278
75% Q3	1.20602	95%	1.22986
50% MED	1.18824	90%	1.22019
25% Q1	1.17206	10%	1.15966
0% MIN	1.1224	5%	1.15214
		1%	1.1369



Quanto à concordância entre os modelos de classificação, encontrou-se um diagnóstico praticamente igual fornecidos pelo modelo de Kokolakis e o da Dirichlet simétrica para qualquer quadro clínico dos pacientes, dado pelos pares de variáveis.

Sob utilização do par (X_1, X_3) na situação de independência, vemos também na tabela 10 uma discordância maior entre os modelos de Kokolakis e o de Independência na classificação dos pacientes, sob baixas prevalências (0.10 e 0.25).

Tabela 10: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_1, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_1, por incidência e tamanho amostral.

a: $\pi_1 = 0.10$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1990	1997	1996
	Pacientes	49876	124900	179819
K-I	Concordância	1576	1636	1660
	Pacientes	44458	112679	163341

b: $\pi_1 = 0.25$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	2000	2000	2000
	Pacientes	50000	125000	180000
K-I	Concordância	1551	1507	1503
	Pacientes	42736	105011	151274

c: π_1=0.50

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1995	2000	2000
	Pacientes	49929	125000	180000
K-I	Concordância	1935	1984	1990
	Pacientes	48852	124273	179353

A diferença entre o diagnóstico emitido pelo modelo de Kokolakis e o de Independência ao utilizar o par (X_1, X_3) se acentua mais sob a prevalência 0.10, quando estas variáveis passam a ter uma pequena covariância nas classes.

Tabela 11: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_1, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_2, por prevalência e tamanho amostral.

a: π_1=0.10

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1999	1999	2000
	Pacientes	49989	124963	180000
K-I	Concordância	1302	1176	1138
	Pacientes	40044	96279	137094

b: π_1=0.25

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1996	1998	1999
	Pacientes	49960	124951	179955
K-I	Concordância	1590	1557	1533
	Pacientes	45306	112576	161253

c: π_1=0.50

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1998	1999	2000
	Pacientes	49970	124966	180000
K-I	Concordância	1937	1985	1998
	Pacientes	49244	124539	179910

As tabelas 12 e 13 que se seguem apresentam outros níveis de concordância entre os modelos ao utilizar o par de indicadores já citado acima. A primeira delas, referente à estrutura e_3, encontra valores praticamente iguais aos encontrados para a estrutura e_4, não apresentados aqui.

Enquanto isso, a tabela 13 apresenta resultados praticamente idênticos aqueles encontrados para as estruturas e_6 a e_9, sob utilização de (X_1, X_3).

Tabela 12: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_1, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_3, por prevalência e tamanho amostral.

a: π_1=0.10

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1995	2000	2000
	Pacientes	49960	125000	180000
K-I	Concordância	1748	1741	1704
	Pacientes	46009	113306	160353

b: π_1=0.25

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1999	2000	2000
	Pacientes	49989	125000	180000
K-I	Concordância	1510	1500	1500
	Pacientes	40920	101968	146965

c: π_1=0.50

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1998	2000	1993
	Pacientes	49965	125000	179527
K-I	Concordância	1792	1799	1831
	Pacientes	46146	115564	168690

Tabela 13: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_1, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_5, por prevalência e tamanho amostral.

a: $\pi_1 = 0.10$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1996	1994	1999
	Pacientes	49954	124823	179957
K-I	Concordância	1573	1620	1660
	Pacientes	44386	112083	163127

b: $\pi_1 = 0.25$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1996	2000	2000
	Pacientes	49941	125000	180000
K-I	Concordância	1522	1510	1501
	Pacientes	42762	105058	150923

c: $\pi_1 = 0.50$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1996	2000	2000
	Pacientes	49929	125000	180000
K-I	Concordância	1930	1979	1991
	Pacientes	48734	124030	179432

Ao utilizarmos o par (X_1, X_2) para classificação, os resultados da concordância mostraram-se praticamente iguais para todas as nove estruturas. Vemos na tabela 14 o grande número de vezes em que o modelo de Kokolakis e o de Independência estabeleceram diagnóstico diferente a partir da mesma evidência, sob baixa prevalência da doença. Para prevalência 0.50 o modelo de Kokolakis estabeleceu regra igual ao modelo de Dirichlet Simétrica e o de Independência em todas as 2000 vezes, sob qualquer estrutura, para qualquer tamanho amostral.

Tabela 14: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_1, X_2) e o número de pacientes com igual classificação pelos modelos sob estrutura e_1, por prevalência e tamanho amostral

a: $\pi_1 = 0.10$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1998	2000	1999
	Pacientes	49976	125000	179962
K-I	Concordância	1142	1030	1007
	Pacientes	38452	92238	131493

b: $\pi_1 = 0.25$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1996	2000	2000
	Pacientes	49928	125000	180000
K-I	Concordância	1817	1892	1943
	Pacientes	46501	119966	176139

c: π_1=0.50

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	2000	2000	2000
	Pacientes	50000	125000	180000
K-I	Concordância	2000	2000	2000
	Pacientes	50000	125000	180000

Quanto ao par (X_2, X_3) surgem duas situações relativamente distintas a respeito da concordância do diagnóstico feito pelos modelos. A primeira, vista na tabela 15, mostra uma menor concordância entre o modelo de Kokolakis e o de Independência para prevalência 0.25, ocorrendo o mesmo nas estruturas e_2, e_3 e e_4.

Tabela 15: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_2, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_1, por prevalência e tamanho amostral

a: π_1=0.10

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1999	2000	2000
	Pacientes	49989	125000	180000
K-I	Concordância	1994	2000	2000
	Pacientes	49925	112083	163127

b: π_1=0.25

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1983	1996	1997
	Pacientes	49669	124794	179776
K-I	Concordância	1829	1845	1839
	Pacientes	46553	116950	168104

c: π_1=0.50

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	2000	2000	2000
	Pacientes	50000	125000	180000
K-I	Concordância	1949	1991	2000
	Pacientes	48759	124480	180000

Lembramos que em todas as quatro primeiras estruturas o par (X_2, X_3) nunca se constituiu no melhor, sob o critério da taxa ótima de erro.

Entretanto, quando as covariâncias assumidas diminuem sua taxa de erro, chegando a torná-lo o melhor par para classificação sob prevalência 0.10, muda a concordância entre os modelos de Kokolakis e o de Independência. Podemos ver na tabela abaixo os resultados para a estrutura e_7, análogos aos das demais (e_6, e_8 e e_9).

Tabela 16: Concordância no diagnóstico entre o modelo de Kokolakis e os demais através de (X_2, X_3) e o número de pacientes com igual classificação pelos modelos sob estrutura e_7, por prevalência e tamanho amostral.

a: $\pi_1 = 0.10$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1974	1989	1998
	Pacientes	49770	124769	179929
K-I	Concordância	1792	1679	1647
	Pacientes	48194	118146	169194

b: $\pi_1 = 0.25$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1999	2000	2000
	Pacientes	49986	125000	180000
K-I	Concordância	1880	1948	1981
	Pacientes	47719	122384	178493

c: $\pi_1 = 0.50$

Modelos		m = 100	m = 250	m = 360
K-DS	Concordância	1981	1987	1990
	Pacientes	49525	124204	179174
K-I	Concordância	1486	1440	1418
	Pacientes	37546	91171	128955

A partir da discordância de diagnóstico verificada entre os modelos de Kokolakis e o de Independência, baseado nos indicadores observados no paciente, estudou-se o comportamento do tipo de erro associado a cada modelo. Até citação em contrário, os resultados a seguir são referentes ao desempenho dos modelos na *amostra atual*.

Percebeu-se, então, que para todas as situações sob utilização de (X_1, X_3) o modelo de Kokolakis teve melhor desempenho quanto ao erro de *falso positivo*, apresentando erro não superior aos demais modelos em quase todas as amostras geradas.

Em compensação, o modelo de Kokolakis apresentou maior *falso negativo* em quase todas as simulações, o que é mostrado na tabela abaixo, referente a estrutura de independência entre os indicadores.

Tabela 17: Número de amostras com erro não superior aos demais modelos sob utilização de (X_1, X_3) para estrutura e_1, por modelo, tipo de erro, prevalência e tamanho amostral.

		Π_1=0.10			Π_1=0.25			Π_1=0.50		
Modelos		100	250	360	100	250	360	100	250	360
FP	K	500	500	500	500	500	500	495	499	500
	I	155	193	202	51	7	3	437	485	490
FN	K	158	193	202	51	7	3	436	485	490
	I	500	500	500	500	500	500	494	499	500

Como vimos para as tabelas anteriores, para prevalência 0.25 havia uma baixa concordância entre os diagnósticos emitidos pelos modelos de Kokolakis e o de Independência. Na tabela 17 vemos agora o sentido da conseqüência originada pela discordância entre as regras.

O resultado acima é praticamente análogo a todas as estruturas estudadas, com exceção das estruturas e_3 e e_4, que apresentam níveis diferentes de discordância para os erros cometidos, embora o sentido ainda seja o mesmo, como mostra a tabela 18.

Tabela 18: Número de amostras com erro não superior aos demais modelos sob utilização de (X_1, X_3) para estrutura e_3, por modelo, tipo de erro, prevalência e tamanho amostral.

		Π_1=0.10			Π_1=0.25			Π_1=0.50		
Modelos		100	250	360	100	250	360	100	250	360
FP	K	500	500	500	499	500	500	498	500	493
	I	253	241	204	9	0	0	290	299	324
FN	K	256	241	204	10	0	0	292	299	331
	I	500	500	500	500	500	500	500	500	500

Entretanto, embora o modelo de Kokolakis tenha se mostrado inferior em relação ao número de amostras onde cometeu maior *falso negativo*, a análise dos resultados da *taxa média de erro* observada permite conclusão diferente.

Em todas as estruturas e prevalências estudadas o modelo de Kokolakis sempre apresentou menor *taxa média de erro* atual, inclusive no caso de independência. Embora já tenhamos afirmado que ele comete maior *falso negativo* na maioria das amostras, seu erro de classificação está bem próximo ao que seria esperado ao aplicarmos a *regra ótima de classificação de Bayes*. A única exceção é para as estruturas e_1, e_5, e_6, e_7, e_8 e e_9 sob prevalência 0.10, quando o *falso negativo* apresentado pelo modelo I (de Kokolakis) é relativamente maior que o erro esperado.

Contudo, como mostra a tabela 19, o modelo I sempre é superior ao III, obtendo níveis maiores de classificação correta, mesmo com o comportamento de cometer maior falso negativo na maioria das amostras.

Tabela 19: Taxa média de erro observada sob utilização de (X_1, X_3) para a estrutura e_1, por tipo de erro, modelo, prevalência e tamanho amostral.

a: π_1=0.10

Erro	Modelo	*F*	m 100	250	360
FP	Kokolakis	0.054	0.0211	0.0333	0.0378
	Independência		0.0929	0.1024	0.1045
FN	Kokolakis	0.0325	0.0618	0.0513	0.0481
	Independência		0.0227	0.0219	0.0223

b: π_1=0.25

Erro	Modelo	*F*	m 100	250	360
FP	Kokolakis	0.045	0.0505	0.0463	0.0454
	Independência		0.1467	0.1507	0.1496
FN	Kokolakis	0.08925	0.0749	0.0805	0.0806
	Independência		0.0258	0.0250	0.0252

c: π_1=0.50

Erro	Modelo	*F*	m 100	250	360
FP	Kokolakis	0.10	0.0876	0.0966	0.0987
	Independência		0.0993	0.0994	0.1005
FN	Kokolakis	0.05	0.0577	0.0523	0.0513
	Independência		0.0486	0.0498	0.0496

Outro resultado, visto acima para Π_1=0.50, foi constante para todas as estruturas, isto é, sob esta prevalência sempre houve equilíbrio entre os modelos de Kokolakis e o de Independência ao se utilizar o par (X_1, X_3) para classificação.

Em relação ao par (X_1, X_2), o comportamento das taxas de erro na comparação entre os modelos mostrou-se igual ao encontrado sob utilização do par (X_1, X_3).

Quanto ao par (X_2, X_3), que sob baixa prevalência (Π_1=0.10) se constituiu no melhor nas últimas quatro estruturas(e_6 a e_9), seu comportamento é inverso ao de (X_1, X_3) nas taxas médias de erro observadas.

As tabelas seguintes, com resultados praticamente iguais para todas as quatro estruturas citadas, mostram que agora o modelo de Kokolakis apresenta maior número de amostras com menor *falso negativo* quando Π_1=0.10.

Tabela 20: Número de amostras com erro não superior aos demais modelos sob utilização de (X_2, X_3) para a estrutura e_6, por tipo de erro, modelo, prevalência e tamanho amostral.

		Π_1=0.10			Π_1=0.25			Π_1=0.50		
Modelos		100	250	360	100	250	360	100	250	360
FP	K	272	111	81	463	481	492	289	248	225
	I	499	500	500	261	254	235	359	376	392
FN	K	471	493	498	326	290	275	412	414	415
	I	241	104	79	414	446	455	228	208	199

Tabela 21: Taxa média de erro observado sob utilização de (X_2, X_3) para a estrutura e_6, por tipo de erro, modelo, prevalência e tamanho amostral.

a: π_1=0.10

			m		
Erro	Modelo	*F*	100	250	360
FP	Kokolakis	0.0234	0.0118	0.0194	0.0212
	Independência		0.0023	0.0014	0.0015
FN	Kokolakis	0.05625	0.0663	0.0602	0.0577
	Independência		0.0881	0.0945	0.0945

b: π_1=0.25

			m		
Erro	Modelo	*F*	100	250	360
FP	Kokolakis	0.0195	0.0200	0.0190	0.0191
	Independência		0.0981	0.1085	0.1194
FN	Kokolakis	0.14063	0.1397	0.1410	0.1402
	Independência		0.1245	0.1156	0.1077

c: π_1=0.50

			m		
Erro	Modelo	*F*	100	250	360
FP	Kokolakis	0.253	0.1627	0.1800	0.1904
	Independência		0.1503	0.1500	0.1496
FN	Kokolakis	0.00625	0.0784	0.0698	0.0626
	Independência		0.1242	0.1243	0.1257

Quanto ao desempenho dos modelos na *amostra futura*, isto é, as conseqüências da aplicação na próxima amostra da regra construída na

atual, percebeu-se igual comportamento ao descrito acima para todas as situações.

Embora, como já esperado, todas as taxas de erro fossem ligeiramente maiores para todos os modelos sob utilização de qualquer par de variáveis, os resultados foram similares aos já discutidos quanto à comparação dos modelos.

6. CONCLUSÕES

Os resultados encontrados para as estruturas simuladas permitiram a distinção de duas situações quanto à detecção do par (X_1, X_3) pelo coeficiente *R* de Kokolakis.

Nas primeiras cinco estruturas (e_1 a e_5), onde sua taxa ótima de erro é sempre menor ou igual a dos outros pares, o desempenho do *R* é influenciado pelo tamanho amostral e pela priori atribuída às classes D_1 e D_2. Sob prevalência 0.10, onde a taxa de erro de (X_1, X_3) é sempre menor com exceção de e_2, *R* só costuma ganhar dos outros coeficientes sob priori 0.10, perdendo ao super-estimar a prevalência com priori 0.50.

Sob prevalência 0.25, onde sua taxa de erro é sempre menor, o *R* é melhor sempre, com desempenho de detecção afetado apenas em amostras pequenas (m=100) sob priori 0.50. Na última prevalência, 0.50, onde ocorre empate deste par com (X_1, X_2) com exceção de e_4, o *R* também é melhor sempre, com problema apenas no desempenho em amostras pequenas (m=100) sob priori 0.10.

Nas últimas quatro estruturas simuladas, as altas covariâncias colocadas entre X_2 e X_3 tornaram a taxa ótima de erro deste par como a menor, seguida pela de (X_1, X_3), sob prevalência 0.10. Sob prevalência 0.25, o par (X_1, X_3) teve taxa ótima de erro sempre menor a dos outros pares e sob prevalência 0.50, houve empate com a taxa do par (X_1, X_2).

Nestes casos, embora o par (X_1, X_3) não se constitua no melhor sob prevalência 0.10, sua detecção é muito alta pelo coeficiente de Kokolakis, que só reduz a indicação deste par sob priori 0.50.

Sob as duas últimas prevalências, quando de fato o par (X_1, X_3) se constitui no melhor, o coeficiente *R* detecta este par com vantagem nítida sobre os demais critérios de seleção, principalmente quando a amostra é pequena (100 pacientes), sob qualquer priori.

Aparentemente, quando a covariância entre as variáveis não assume um valor expressivo de forma a distanciar-se da situação de independência geral entre os indicadores, o coeficiente de Kokolakis é melhor que os demais para detectar as melhores variáveis.

Entretanto, seu desempenho sob baixas prevalências pode depender essencialmente da priori utilizada, que não deve super-estimar a prevalência da doença. Sob demais prevalências, desde que se trabalhe com amostras grandes (250 ou 360 pacientes), a priori especificada não afeta seu desempenho na detecção do melhor par de variáveis.

Quando duas variáveis apresentam covariâncias expressivas entre si nas duas classes D_1 e D_2, o coeficiente R é muito sensível para identificar o melhor par de variáveis sob prevalências 0.25 e 0.50, convivendo bem com a concorrência dos demais pares quanto à taxa ótima de erro.

Contudo, o problema apresentado sob baixa prevalência em situações próximas à independência volta a ocorrer só que de forma diferente. Embora o par (X_1, X_3) apresente maior taxa ótima e atual de erro, superiores às do par (X_1, X_2), o coeficiente de Kokolakis tem a tendência de detecção excessiva do par fixado, composto pelas melhores variáveis individuais.

A partir da detecção do par (X_1, X_3), a regra de classificação criada consegue um bom desempenho ao utilizar o modelo de Kokolakis, embora sem vantagem expressiva sobre o modelo dado pela Dirichlet Simétrica e o modelo de Independência.

Kokolakis(1983), ao estudar a classificação de pacientes com problemas cardíacos sob prevalência de 26% dos casos, encontrou este mesmo resultado de classificação correta, sob a utilização de duas variáveis entre cinco disponíveis. Entretanto, a medida que seu modelo inclui mais variáveis estabelece-se sua superioridade, mesmo com pequeno tamanho amostral (não excedendo 100 pacientes). Para um grande número de variáveis, o autor encontrou melhor desempenho do modelo de Independência sobre o modelo da Dirichlet Simétrica quando o tamanho amostral é pequeno. Isto talvez possa ser explicado devido a priori Dirichlet não tratar bem o problema de caselas com freqüências nulas,

enquanto o modelo que assume independência dos indicadores não é realista e tem efeitos piores quando o tamanho amostral é grande.

Voltando ao nosso estudo, ainda em relação ao desempenho dos modelos de classificação sob utilização de um par de variáveis, ocorrem novamente duas situações interessantes.

Como vimos, sob utilização do par (X_1,X_3) o modelo de classificação de Kokolakis sempre apresenta menor erro médio atual, mas apresenta especificamente um maior número de amostras onde se apresenta como o modelo que comete maior taxa de erro de *falso negativo*.

Dependendo do tipo de enfermidade a ser diagnosticada, para providenciar um tratamento urgente, este comportamento pode ter conseqüências ruins. Entretanto, o erro atual do *falso negativo* cometido apresenta-se bem próximo ao erro ótimo esperado sob utilização da regra de Bayes, ou seja, este tipo de erro cometido por Kokolakis não decepciona seu desempenho.

Contudo, sob prevalência 0.10 em estruturas onde o par (X_1,X_3) é sobrepujado por (X_2,X_3) quanto à taxa ótima de erro, a seleção do primeiro par pode ter uma conseqüência pior. Neste caso, o erro médio atual de *falso negativo* é expressivamente maior do que o cometido pelo modelo de Independência e o erro ótimo esperado, principalmente se o tamanho amostral for pequeno. Lembramos a conseqüência danosa deste fato, devido à grande detecção do par (X_1,X_3) pelo coeficiente R, observada nesta situação citada.

Se o par (X_2,X_3), efetivamente o melhor, fosse selecionado haveria um maior número de amostras com menor *falso negativo*, com taxa atual ligeiramente acima da esperada.

De certa forma, comparando a detecção do par (X_1,X_3) e o desempenho da classificação pelo modelo de Kokolakis sob sua utilização, parece que ambas são muito afetadas pelo comportamento individual dos testes em cada classe D_1 e D_2, ou seja, por sua *sensibilidade* e sua *especificidade*.

Intencionalmente simulamos estruturas de testes com maior *sensibilidade* do que *especificidade*, e simultaneamente um teste é

melhor do que outro em relação a qualquer uma destas qualidades.

Assim, a observação "positiva" do indicador permite maior segurança para emitir diagnóstico "positivo", isto é, "paciente pertence à classe D_1", do que a correspondente alocação resultante de observação "negativa". Neste caso, não sendo observada a característica desejada no indicador clínico, devido a menor *especificidade* o modelo de classificação não teria a mesma ousadia para emitir diagnóstico, agora "paciente pertence à classe D_2". Logo, explica-se o maior *falso positivo* de Kokolakis, acarretado por utilizar "em excesso" a *sensibilidade* alta dos testes e o maior *falso negativo*, decorrente de sua relutância em classificar o paciente como doente, devido à *especificidade* menor dos testes.

Devido a esta característica aparente de valorizar demais a distribuição marginal dos testes e escolher em excesso a composição destas variáveis que conjuntamente não teria menor taxa de erro, poder-se-ia pensar numa seleção sequencial dos indicadores.

Teríamos então, a partir de um grupo já selecionado de variáveis com as quais se trabalha, a seleção de um novo indicador baseado no maior valor assumido pelo coeficiente *R*. Mesmo que este novo sub-grupo de variáveis não fosse ótimo, vale lembrar que se o mesmo for basicamente constituído pelas melhores variáveis individuais, consegue-se conviver razoavelmente bem no caso de perda de observação de um dos indicadores.

O critério que seria estabelecido para inclusão sequencial de testes poderia depender da distribuição do coeficiente *R*, a ser melhor estudada para várias estruturas diferentes de dependência entre as variáveis.

Entretanto, vemos com preocupação a necessidade de um grande banco de dados para estabelecer a seleção e testar a regra nos pacientes, para uma grande dimensionalidade de indicadores, que podem geralmente estarem disponíveis em problemas reais de diagnóstico médico.

Embora tenhamos estudado a aplicação das regras na *amostra futura*, não tínhamos intenção de estabelecer conclusões a respeito da *transferibilidade* dos modelos de classificação.

Isto poderia ser feito estabelecendo-se diferentes critérios amostrais de centros clínicos, onde a prevalência da doença, *sensibilidade/especificidade* dos testes e sua estrutura de dependência poderiam ser variáveis, seguindo uma distribuição fixada.

Além disso, também não estudamos o desempenho do coeficiente e modelo de classificação de Kokolakis para uma classe mais numerosa de doenças, onde algumas poderiam ocorrer simultaneamente em um mesmo paciente. No caso onde $D=\{D_1, D_2, D_1 \cup D_2, D_3\}$, onde D_1 e D_2 seriam duas doenças distintas e D_3 a não-ocorrência de nenhuma destas, poderíamos talvez seguir a idéia de Pereira e Pericchi(1985), que sugerem uma discriminação inicial entre $(D_1 \cup D_2)$ e D_3, para uma posterior discriminação entre D_1 e D_2, a partir da eficiência em separar os primeiros grupos.

Desde que a situação permita no problema estudado, seria interessante também buscar a melhor categorização binária, para o coeficiente *R* melhorar sua detecção e a regra construída apresentar melhor desempenho na alocação. Contudo, isto só deveria ser feito se houver facilidade em observar os indicadores sob a nova categorização nos mais diferentes centros clínicos que tratam a enfermidade sob estudo.

De uma maneira geral, poderíamos dizer que o coeficiente *R* de Kokolakis é um bom critério para seleção de variáveis, conseguindo também melhores taxas de erro do que os outros modelos, sob utilização dos indicadores já escolhidos.

Entretanto, o critério que derivou este coeficiente deve ser compreendido estritamente sob a utilização da priori que tenta responder aos requisitos de Lindley(1978). Assim, na verdade, não teríamos exatamente um critério mas sim um coeficiente que pode ser criado a partir do emprego desta priori, que usa misturas de prioris

Dirichlet.

Além disso, seria interessante um estudo de seu desempenho num contexto onde a conseqüência dos possíveis erros cometidos não assumisse uma estrutura tão simples, tratando-os com mesmo peso, mas sim sob o enfoque de Teoria da Decisão.

RESUMO

Com ênfase especial ao diagnóstico médico, para selecionar o melhor sub-grupo de k testes clínicos binários entre d disponíveis e classificar o paciente como portador(D_1) ou não(D_2) de uma doença a partir de sua observação, é utilizado o critério de Kokolakis. Este critério, sob enfoque Bayesiano, é baseado nos pesos *a posteriori* de uma mistura de $(d)_k$ prioris, cada uma correspondendo a uma particular ordenação das k variáveis, invariante sob permutação dos testes. Além de permitir correlação positiva entre probabilidades de caselas associadas à seqüências similares, esta nova caracterização de prioris implica que *a priori* cada teste é sucessivamente menos útil do que o anterior.

O estudo analisa o caso especial onde d=3 e k=2, para nove diferentes estruturas de dependência entre os três testes clínicos, três níveis de prevalência da doença e três níveis de probabilidade a priori especificada para D_1. Através de simulações, onde os três testes são definidos com *sensibilidade* e *especificidade* conhecidas, é comparado o poder de detecção do par ótimo de variáveis pelo critério proposto e pelos critérios de Akaike, Bhattacharyya, Jeffreys e Lindley.

A partir do par selecionado de testes, a regra de classificação derivada é comparada com as dos modelos de Independência e de Dirichlet Simétrica em relação à taxa de erro, encontrando-se um desempenho relativamente melhor do modelo de Kokolakis sobre os demais.

REFERÊNCIAS BIBLIOGRÁFICAS

AITCHISON, J. and LAUDER, I.J.(1979). Statistical diagnosis from imprecise data. *Biometrika*, 66, 3, 475-483.

ANDERSON, J.A.(1972). Separate sample logistic discrimination. *Biometrika*, 59, 1, 19-35.

BERNARDO, J.M. and BERMUDEZ, J.D.(1985). The choice of variables in probabilistc classification (with discussion). In *Bayesian Statistics* 2 (J.M.Bernardo, M.H.DeGroot, D.V.Lindley, A.F.M.Smith, eds), pp.67-82. Amsterdam, North-Holland.

BHATTACHARYYA, A.(1946). On a measure of divergence between two multinomial populations. *Sankhyã*, 7, 401-406.

BROWN, P.J.(1976). Remarks on some statistical methods for medical diagnosis. *J. R. Statist. Soc. A*, 139, Part 1, 104-107.

BURBEA, J. and RAO, C.R.(1982). On the convexity of some divergence measures based on entropy functions. IEEE *Transactions on Information Theory*, vol IT-28, 3, 489-495.

COCHRAN, W.G.(1964). On the performance of the linear discriminant function. *Technometrics*, 6, 179-190.

COCHRAN, W.G. and HOPKINS, C.E.(1961). Some classification problems with multivariate qualitative data. *Biometrics*, 17, 10-32.

COVER, T.M.(1974). The best two independent measurements are not the two best. IEEE *Trans. Syst., Man, Cybern.* SMC-4, 116-7.

COX, D.R.(1966). Some procedures associated with the logistic qualitative response curve. In *Research Papers in Statistics: Festschrift for J. Neyman*, ed. F.N.David, pp.55-71, New York: Wiley.

DAY, N.E. and KERRIDGE, D.F.(1967 . A general maximum likelihood discriminant. *Biometrics* 23, 313-323

DAWID, A.P.(1976). Properties of diagnostic data distributions. *Biometrics*, 32, 647-658.

DE FINETTI, B.(1965). Methods for discriminating levels of partial knowledge concerning a test item. *British Journal of Mathematical and Statistical Psychology*, 18, Part 1, 87-123.

ELASHOFF, J.D., R.M. ELASHOFF and G.E. GOLDMAN(1967). On the choice of variables in classification problem with dichotomous variables. *Biometrika*, 54, 668-670.

FARVER, T.B. and DUNN, O.J.(1979). Stepwise variable selection in classification problems. *Biom. J.* **21**, 2, 145-153.

FUSIKOSHI, Y.(1985). Selection of variables in two-group discriminant analysis by error rate and Akaike's information criteria. *Journal of Multivariate Analysis* **17**, 27-37.

GOLDSTEIN, M. and DILLON, W.R.(1977). A stepwise discrete variable selection procedure. *Commun. Statist. - Theor. Meth.*, **A6**(14), 1423-1436.

GOOD, I.J.(1965). *The Estimation of Probabilities.* MIT Press, Cambridge, Massachusetts.

____ (1985). Weight of evidence: a brief survey (with discussion). In *Bayesian Statistics* 2 (J.M.Bernardo, M.H.DeGroot, D.V.Lindley, A.F.M.Smith, eds), pp.67-82. Amsterdam, North-Holland.

GOOD, I.J. and CARD, W.I.(1971). The logistic process with special reference to errors. *Methods of Information in Medicine*, **10**, 176-188.

GOWER, J.C. and PAYNE, R.W.(1975). A comparision of different criteria for selecting binary tests in diagnostic keys. *Biometrika*, **62**, 3, 665-672.

HAERTING, J.(1983). Special properties in selection performance of qualitative variables in discriminant analysis. *Biom. J.*, **25**, 215-222.

HUGHES, G.F.(1968). On the mean accuracy of statistical pattern recognizers. IEEE *Trans. Information Theory*, **14**, 55-63.

KOKOLAKIS, G.E.(1983). A new look at the problem of classification with binary data. *The Statistician*, **32**, 144-152.

____ (1985). A Bayesian criterion for the selection of binary features in classification problems. In *Bayesian Statistics* 2 (J.M.Bernardo, M.H.DeGroot, D.V.Lindley, A.F.M.Smith, eds), pp.673-680.

LACHIN, J.M.(1973). On a stepwise procedure for two population Bayes decision rules using discrete variables. *Biometrics*, **29**, 551-564.

LINDLEY, D.V.(1978). The Bayesian Approach. *Scandinavian Journal of Statistics*, **5**, 1-26.

MATUSITA, K.(1951). On the theory of statistical decision functions. *Ann. Inst. Stat. Math.*, **3**, 17-35.

MOORE II, D.H.(1973). Evaluation of five discrimination procedures for binary variables. *JASA*, **68**, 342, 399-404.

MORAN, P.A.P.(1985). Parsimony in the construction of diagnostic scales. *British Journal of Mathematical and Statistical Psychology*, **38**, 202-205.

NOVICK, M.R. and GRIZZLE, J.E.(1965). A Bayesian approach to the analysis of data from clinical trials. *JASA*, **60**, 81-96.

PEREIRA, C.A. de B.(1989). Influence diagrams and medical diagnosis. To appear.

PEREIRA, C.A. de B. and BARLOW, R.E.(1989). Medical diagnosis using influence diagrams. To appear.

PEREIRA, C.A. de B. and PERICCHI, L.R.(1988). Analysis of diagnosability. *J. R. Statist Soc. C*, (submitted).

PEREIRA, C.A. de B. and VIANA, M.A.G.(1982). *Elementos de Inferência Bayesiana*. V SINAPE, IME-USP, São Paulo.

SAKAMOTO, Y. and AKAIKE, H.(1978). Analysis of cross classified data by AIC. *Ann. Inst. Statist. Math.*, **30**, Part B, 185-197.

SAKAMOTO, Y., ISHIGURO, M. and KITAGAWA, G.(1983). *Akaike Information Criterion Statistics*. Tokyo, KTK Scientific Publishers.

SAVAGE, L.J.(1971). Elicitation of personal probabilities and expectations. *JASA*, **66**, 336, 783-801.

SCHMITZ, P.I.M. and HABBEMA, J.D.F.(1985). A simulation study of the performance of five discriminant analysis methods for mixtures of continuous and binary variables. *J. Statist. Comput. Simul.*, **23**, 69-95.

SKARABIS, H.(1970). Mathematische grundlagen und praktische aspekte der diskrimination und klassifikation. Würzburg:Physica.

SPIEGELHALTER, D.J. and KNILL-JONES, R.P.(1984). Statistical and knowledge-based approaches to clinical decision-support systems, with an application in gastroenterology (with discussion). *J. R. Statist. Soc. A*, **147**, Part 1, 35-77.

STANISH, W.M. and ALLRED, R.U.(1981). Categorical variable selection based on entropy reduction. *Commun. Statist. - Theor.Meth.* **A10**(17), 1733-1750.

TEATHER, D.(1974). Statistical techniques for diagnosis. *J. R. Statist. Soc.* A, **137**, Part 2, 231-244.

THIBODEAU, L.A.(1981). Evaluating diagnostic tests. *Biometrics*, **37**, 801-804.

TITTERINGTON, D.M., MURRAY, G.D., MURRAY, L.S., SPIEGELHALTER, D.J., SKENE, A.M., HABBEMA, J.D.F. and GELPKE, G.J.(1981). Comparision of discrimination techniques applied to a complex data set of head injured patients (with discussion). *J. R. Statist. Soc.* A, **144**, Part 2, 145-175.

TOUSSAINT, G.T.(1971). Note on optimal selection of independent binary features for pattern recognition. IEEE *Trans. Inform. Theory*, IT-17, 618.

VACEK, P.M.(1985). The effect of conditional dependence on the evaluation of diagnostic tests. *Biometrics*, **41**, 959-968.

WERNECKE, K.-D., UNGER, S. and KALB, G.(1986). The use of combined classifiers in medical functional diagnostic. *Biom. J.*,28,1, 81-88.