PROVAS ACADÉMICAS
NA FACULDADE DE MEDICINA DA UNIVERSIDADE DE LISBOA
INSTITUTO DE FORMAÇÃO AVANÇADA

Doutoramento:
Medicina

Nome do Aluno:
Isabel Maria França Dória

Tema da Tese:
Representações euclidianas de dados - uma abordagem para variáveis heterogéneas

Área:
Medicina

Especialidade:
Biomatemática

Data da Defesa:
05/05/2009

Classificação:
Aprovada com Distinção e Louvor por Unanimidade

Júri:
Presidiu o Presidente do Conselho Científico da FMUL, Professor Doutor Henrique Bicha Castelo e estiveram presentes os vogais: Professores Doutores Georges le Calvé, da Universidade de Reines 2, França, Paulo Jorge Mota de Pinho Gomes, da Universidade Nova de Lisboa, José Manuel Pereira Miguel, Maria Luísa Figueira, Maria Helena Nicolau, Miguel Oliveira da Silva e António Vaz Carneiro, todos da Universidade de Lisboa.

RESUMO

Esta dissertação insere-se na área de Análise de Dados Multivariados, sob o tema de Representações de Dados. Os objectivos da tese são de ordem metodológica, incluindo desenvolvimento de *software e* aplicação a dados reais no âmbito da Biomatemática. O principal objectivo consiste na representação simultânea de variáveis de diferentes tipos ou seja de variáveis heterogéneas. Para este efeito foram usados coeficientes existentes na literatura, alguns dos quais ainda pouco estudados.

Em Biomatemática, como em outras disciplinas, as variáveis descritoras dos indivíduos são frequentemente de natureza diversa - esta situação é habitual, por exemplo, em investigação baseada em inquéritos e questionários. Contudo, a heterogeneidade das variáveis cria problemas matemáticos e estatísticos específicos, que são difíceis de resolver quando se pretendem obter representações euclidianas. A abordagem apresentada neste trabalho dá uma contribuição para a representação de variáveis heterogéneas.

Em geral, os dados em estudo são considerados sob a forma de uma matriz "numérica" que cruza um conjunto de indivíduos e um conjunto de variáveis. Esta matriz é transformada numa matriz de semelhanças ou de dissemelhanças. É esta matriz de proximidades que vai ser representada. Esta abordagem, mais geral do que a abordagem tradicional, permite tratar de uma maneira unificada os diversos métodos factoriais (Análise em Componentes Principais (ACP), a Análise Factorial das Correspondências e outros), por um lado, e por outro, os de Posicionamento Multidimensional *(Multidimensional Scaling).*

Tratamos também de casos em que não existe representação euclidiana exacta dos dados. Interessamo-nos em particular, pela categoria importante das transformações monótonas, que "perturbem ao mínimo" os dados originais. Mostramos que os métodos "da constante aditiva" não são fiáveis (a constante é, com frequência,
extremamente grande em relação às proximidades iniciais). A transformação pela função potência, por outro lado, é uma proposta que parece ser muito promissora.

São abordados em pormenor os coeficientes s, s_{Lc} e P_L (Le Calvé, 1977) e os coeficientes de afinidade - a, a_w, a_s, VAL_{AW}, VAL_{AS} e os generalizados (e.g., BacelarNicolau, 1980, 1988, 2002; Nicolau *et al.,* 2007) - que permitem tratar simultaneamente variáveis do mesmo ou de diferentes tipos. A abordagem probabilística destes coeficientes - VAL_{AW}, VAL_{AS} e P_L - tem a mesma origem (ver Lerman, 1970). Quando se comparam variáveis do mesmo tipo, alguns daqueles coeficientes coincidem com coeficientes conhecidos, tal como o coeficiente de correlação de Pearson.

Generalizaram-se os coeficientes s, s_{Lc} e P_L para serem utilizados com dados simbólicos (e.g., Doria *et al.,* 2007) e variáveis de ordem sequencial. Verificou-se que, em certos casos, estes permitem comparar variáveis com dados omissos sem recorrer a métodos de imputação e são uma mais valia para o caso das variáveis com categorias parcialmente ordenadas (e.g., Doria *et ai.,* 2006b), assim como na comparação de ultramétricas (e.g., Doria *et ai.,* 2000). Paralelamente, desenvolveu-se *software* para o cálculo dos coeficientes s, s_{Lc} e P_L e aplicaram-se estes coeficientes a problemas reais do domínio da Biomatemática (e.g., Doria *et ai.,* 2006a; Doria *et ai.,* 2007). As análises em componentes principais das matrizes de semelhanças S_{Lc} e P_L conduziram, de forma geral, a boas visualizações das relações entre as variáveis. No caso particular de todas as variáveis serem métricas, a ACP da matriz de semelhanças S_LC coincide com a ACP clássica, a menos de uma translação. Também foi possível visualizar as relações entre as variáveis, através de dendrogramas que resultaram de análises classificatórias hierárquicas ascendentes aplicadas directamente sobre as matrizes de semelhanças S, S_{Lc} e P_L.

Palavras-Chave: variáveis heterogéneas, variáveis simbólicas, coeficiente de semelhança, coeficiente probabilístico, distância, distância euclidiana, análise em componentes principais, análise classificatória hierárquica ascendente.