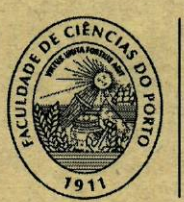
Luís Miguel Almeida da Silva

Selecção de variáveis em *microarrays* de ADN



FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio de 2003

Luís Miguel Almeida da Silva

Selecção de variáveis em *microarrays* de ADN

FACULDADE DE CIÊNCIAS
UNIVERSIDADE DO PORTO

Tese submetida à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Mestre em Estatística

Departamento de Matemática Aplicada
Faculdade de Ciências da Universidade do Porto
Maio 2003

Aos meus pais e à Ana...

Agradecimentos

Nem sempre é fácil expressar como estamos agradecidos a todos aqueles que contribuem para o nosso sucesso. No entanto, não queria deixar de agradecer a colaboração e ajuda demonstradas a todos aqueles que de alguma forma contribuiram directa ou indirectamente para a concretização deste trabalho. Em particular,

ao Doutor Joaquim P. Costa, meu orientador, por toda a disponibilidade e dedicação que sempre mostrou, orientando-me e conduzindo-me ao longo deste trabalho, mesmo nas alturas mais difíceis;

aos meus pais por todo o apoio, carinho e dedicação sempre demonstrados, bem como por terem proporcionado todas as condições para que este trabalho se realizasse;

à Ana, por todo o carinho, ajuda, força e encorajamento que sempre me transmitiu, principalmente nos momentos mais difíceis;

aos meus colegas de mestrado e de faculdade, por todo o apoio e excelentes trocas de ideias ao longo deste trabalho;

a todos o meu muito obrigado.

Resumo

A aplicação de um regime específico de quimioterapia, depende de um correcto diagnóstico do paciente. Actualmente, esse diagnóstico não é efectuado de uma forma sistemática e geral, necessitando da intervenção de diferentes especialistas. Com a monitorização da expressão de milhares de genes em simultâneo, a recente tecnologia dos *microarrays* de ADN representa um grande passo para a sistematização do diagnóstico oncológico. No entanto, esta tecnologia produz informação em que o número de variáveis (genes) excede largamente o número de observações (amostras), o que dificulta a utilização das ferramentas estatísticas habituais de classificação. Neste sentido, torna-se necessário implementar estratégias de redução da dimensão do espaço predictor. Contudo, esta redução, que equivale a seleccionar um subconjunto de genes do conjunto inicial, deve ser extremamente direccionada, pois sabe-se que apesar do seu elevado número, apenas uma pequena parte dos genes monitorizados determina o tipo de tecido. Identificar genes com potencial predictivo é um objectivo central dos estudos de *microarray* aplicados à classificação de tumores. A criação de mecanismos que permitam reter apenas estes genes é essencial, tanto a nível estatístico (pois só com redução de dimensão poderemos aplicar as metodologias habituais) como a nível de interpretação biológica.

O presente estudo pretende efectuar uma introdução a este novo tipo de tecnologia e de dados, procurando mostrar algumas das metodologias utilizadas para a sua análise. Em particular, centramo-nos no estudo de diferentes estratégias de escolha de genes predictivos e na sua utilização para a classificação de tumores cancerígenos.

Abstract

The application of specific chemotherapy regimens depends on the patient's correct diagnosis. There has been no general or sistematic approach to cancer diagnosis. Rather, it depends on the experience of several specialists. With the ability to measure the expression of thousands of genes simultaneously, DNA microarrays represent a great step for cancer diagnosis. However, this ability to measure gene expression has resulted in data with the number of variables (genes) far exceeding the number of samples; and standard statistical tools in classification do not work well. Strategies to reduce dimension have to be implemented. While the number of measured genes is in the thousands, it is assumed that only a few marker genes determine the type of tissue. Thus, a main objective in microarray studies (applied to cancer classification) is to identify those genes with predictive potential. In this sense, it is essential to create stategies that allow us to retain those genes. Dimension reduction is important in two ways. First, in a statistical level, standard tools can be used; second, in a biological level, interpretation and understanding on how the genome works will be facilitated. This study pretends to make an introduction to this recent technology and data. It is made a kind of survey on different methodologies that are being used in microarray analysis. In particular, we will focus on gene selection and in cancer classification.

Lista de abreviaturas

ACP - análise em componentes principais

ADN - ácido desoxirribonucleico

ALL - leucemia aguda linfocítica

ALL-B - subtipo de ALL de células B

ALL-T - subtipo de ALL de células T

AML - leucemia aguda mielógena

ARN - ácido ribonucleico

B-CLL - leucemia linfocítica crónica de células B

BL - linfoma de Burkitt

cADN - de *cDNA*, *complementary DNA*

CP - componentes principais

DLBCL - linfoma difuso de células B grandes

EWS - sarcoma de Ewing

FL - linfoma folicular

LD - discriminante logístico

LDA - análise discriminante linear

mARN - ARN mensageiro

MQD - média quadrática dentro das populações

MQE - média quadrática entre as populações

NB - neuroblastoma

PLS - *partial least squares* (mínimos quadrados parciais)

PS - *prediction strength*

QDA - análise discriminante quadrática

RMS - rabdomiosarcoma

SOM - *self organizing maps*

SQD - soma dos quadrados dentro das populações

SQE - soma dos quadrados entre as populações

SRBCT - *small round blue cell tumors*

VC - validação cruzada

Conteúdo

Introdução

O desafio do tratamento oncológico tem sido a aplicação de terapias específicas a cada tipo de tumor, maximizando a eficácia e minimizando a toxicidade. Actualmente, a classificação de tumores assenta numa variedade de variáveis morfológicas, clínicas e moleculares. É possível que classes já existentes sejam heterogéneas e compreendam subtipos molecularmente distintos, com percursos clínicos e/ou resposta a tratamentos completamente diferentes. Por exemplo, cancros da próstata morfologicamente idênticos podem ter percursos clínicos completamente diferentes, desde a inactividade ao longo de décadas até ao crescimento explosivo, provocando a rápida morte do paciente. Com a monitorização da expressão de milhares de genes em simultâneo, a recente tecnologia dos *microarrays* de ADN representa um grande passo para a sistematização do diagnóstico oncológico. Técnicas como *microarrays* de cADN ou de oligonucleótidos, podem levar a uma melhor compreensão das variações moleculares entre tumores, conduzindo a uma classificação mais segura.

Podemos considerar três tipos de problemas estatísticos no que concerne a classificação de tumores: a identificação de novas classes, a classificação de novas amostras em classes conhecidas e a identificação de genes "marcadores"que caracterizam as diferentes classes. Soluções para o primeiro problema recorrem a métodos de *análise classificatória* como a *classificação hierárquica* [11, 2] ou métodos de particionamento como os SOM (*self organizing maps*) [31].
Uma particularidade deste tipo de dados é o número largamente superior de variáveis (genes) relativamente ao número de observações (amostras/pacientes), o que dificulta o uso das ferramentas estatísticas habituais de classificação. As soluções para o segundo problema exigem, assim, novas metodologias de classificação ou uma redução efectiva da dimensão (isto é, do número de variáveis) que permita o uso das metodologias já existentes. Contudo, esta redução deverá ser extremamente direccionada pois sabe-se que apenas uma pequena parte dos genes caracteriza os diferentes tumores. Entramos assim no terceiro problema e este é, talvez, o fundamental.

A presente dissertação centra-se precisamente nos dois últimos problemas: a escolha de genes predictivos e a sua utilização para a classificação de novas amostras.
O primeiro capítulo pretende efectuar uma introdução aos *microarrays* de ADN, explorando as suas capacidades, a sua justificação e a sua obtenção, precedida de

uma pequena abordagem conceptual. Metodologias possíveis de serem empregues na análise deste tipo de dados são também descritas, bem como o problema dimensional colocado.

O segundo capítulo é reservado à selecção de genes predictivos, isto é, genes com capacidade discriminativa das diferentes classes. Métodos para o problema de duas classes e problema multi-classes são tratados separadamente.

A redução de dimensão efectuada com a escolha selectiva de genes pode, em alguns casos, não ser suficiente. Assim, o terceiro capítulo dedica-se à descrição de metodologias que permitam complementar as anteriores e até melhorar a performance do classificador. A análise de componentes principais, os mínimos quadrados parciais e as variáveis canónicas são aqui apresentados.

O quarto capítulo é dedicado aos métodos de análise classificatória. O capítulo inicia com uma retrospectiva de algumas aplicações destes métodos a estudos *microarray*. A secção final dedica-se a um método proposto por Dettling *et al.* [6] e que compreende uma mistura de análise discriminante com análise classificatória para encontrar subconjuntos de genes "marcadores"cuja expressão claramente discrimina as diferentes classes.

A descrição dos resultados obtidos pelas diferentes referências e as comparações entre os diferentes métodos é deixada para o quinto capítulo onde são também retratadas algumas experiências por mim realizadas.

Finalmente, as conclusões tiradas ao longo do estudo são reunidas num capítulo final onde também se incluem algumas perspectivas e objectivos futuros nesta área recente.

Capítulo 1

Células e genoma. *Microarrays* de ADN.

1.1 Alguns conceitos

A constituição celular dos seres vivos representa actualmente o conceito fundamental de toda a Biologia. Segundo ele, animais e plantas são constituídos por territórios limitados e habitualmente bem individualizados, cada um dos quais representa uma *célula.*

Cada célula, embora vizinha de outras células e embora com elas estabeleça íntimos contactos de contiguidade, mantém um certo grau de independência, que torna possível o seu crescimento, a sua multiplicação ou a sua morte, independentemente daquelas que a rodeiam. Mas esta independência anatómica e funcional não é completa. Cada célula sofre a influência mecânica daquelas que se encontram à sua volta e pode ser atingida por estímulos excitantes ou paralisantes de células longínquas ou próximas. Deste modo, animais e plantas representam a associação harmónica de milhões de unidades morfológicas e fisiológicas que, embora mantendo grande independência, mutuamente se influenciam.

Existem dois tipos de organismos classificados de acordo com o tipo de células que os constituem: *procariotas* e *eucariotas.* As bactérias são um exemplo de organismos procariotas, enquanto gatos, árvores, humanos e a maior parte dos seres que habitualmente vemos são exemplos de organismos eucariotas. As diferenças entre estes tipos de organismos residem fundamentalmente na estrutura e processos metabólicos das respectivas células. As células procariotas são menores e possuem uma estrutura muito mais simples. Uma das suas principais características é a de não possuir um *núcleo* diferenciado do restante material celular, resultante de não possuirem uma membrana nuclear. Já as células eucariotas possuem um núcleo bem diferenciado (envolto na membrana nuclear), para além de um vasto conjunto de sub-estruturas, também elas bem diferenciadas, cada qual com a sua função no metabolismo celular (respiração, produção de açúcares, etc.). Estas estruturas designam-se de *organelos* e

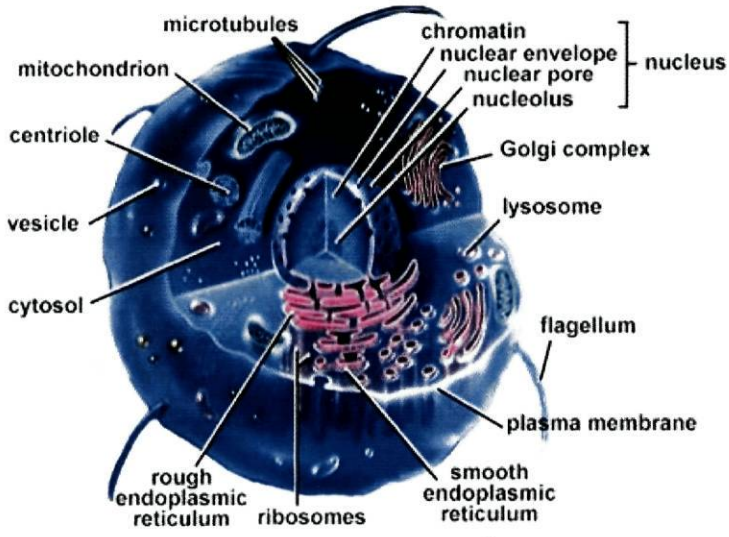


Figura 1.1: Célula eucariota. Pode observar-se o núcleo bem diferenciado, bem como todo um conjunto de sub-estruturas bem diferenciadas - os *organelos*. Figura adaptada de [4].

são exemplos os centríolos, complexos de Golgi, mitocôndria entre outros (ver figura 1.1). Uma das características mais importantes das células é a sua capacidade de crescer num ambiente apropriado e de se multiplicar. A noção de célula e, depois, o estudo cada vez mais profundo da sua constituição morfológica e bioquímica têm tido as mais profundas repercussões no desenrolar dos conhecimentos biológicos, incluindo a patologia e a genética. A genética ocupa-se do estudo de uma estrutura muito importante da célula - o núcleo. É nele que estão contidos os *cromossomas*. Este material, que existe em número constante para cada espécie (23 pares para o ser humano), são o suporte dos *genes*. Um gene consiste numa porção contínua de uma molécula de ADN, a partir da qual, através de uma maquinaria molecular complexa, é lida informação que permite produzir um tipo ou alguns tipos de proteínas (sendo estas, componentes estruturantes das células).

Os *ácidos desoxirribonucleicos* ou ADN, são substâncias altamente específicas das quais depende toda a actividade de ordem morfológica ou metabólica. A molécula de ADN é formada por duas cadeias helicoidais, cada uma das quais, constituída por uma sequência de unidades moleculares básicas denominadas *nucleótidos* (ver figura 1.2). Cada nucleótido é constituído por uma unidade fosfato, uma unidade de desoxirribose e uma de quatro bases azotadas: adenina (A), guanina (G), citosina (C) ou timina (T). As duas hélices unem-se por ligações de hidrogénio entre as suas bases azotadas, mas de tal modo que estas ligações só são possíveis entre T e A e entre C e G. Esta propriedade muito importante do ADN designa-se por *complementaridade*. As mitocôndrias também contêm ADN, mas em quantidade muito reduzida. Em conjunto, ADN dos cromossomas e das mitocôndrias, formam o *genoma* do organismo. Crê-se que toda a informação hereditária de um organismo está codificada neste material genético (para uma introdução mais pormenorizada da biologia molecular veja-se [4]). Há, no entanto, um conjunto de factores (por exemplo radiação) que podem provocar alterações nas moléculas de ADN (e, portanto, nos genes), provocando a morte das

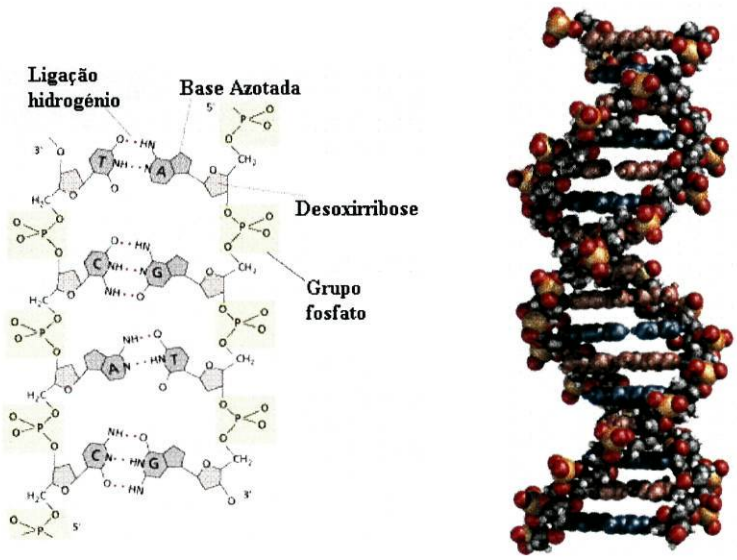


Figura 1.2: À esquerda, esquema da estrutura da molécula de ADN, onde se podem observar os diferentes nucleótidos; à direita, imagem tridimensional da dupla hélice de ADN. Figura adaptada de [4].

células ou a sua proliferação anormal e anárquica através de divisões e diferenciações descontroladas, conduzindo ao aparecimento de tumores cancerígenos.
Nos dias de hoje, o estudo do cancro é central nas áreas da Medicina. A tentativa de entendimento do seu aparecimento, evolução e funcionamento, não é, ainda, perfeitamente conhecido, levando os cientistas a procurar as explicações no suporte de toda a construção e edificação celular que é o ADN, e em particular, os genes.

1.2 *Microarrays* de ADN

Genomics aims to provide biologists with the equivalent of chemistry's Periodic Table.

Eric S. Lander
Array of Hope - Nature Genetics, 21

O maior objectivo da investigação genética é tentar perceber como o genoma funciona. Algumas questões que podem ser colocadas são:

- quais os papéis funcionais que desempenham determinados genes e em que processos celulares participam;
- como são os genes regulados;

- como é que os genes interagem;
- como é que a expressão de um gene muda com doenças ou tratamentos.

A tecnologia dos *microarrays* de ADN surge como um grande passo para responder às questões anteriores, podendo ainda tornar-se como uma ferramenta essencial para a sistematização[1] do diagnóstico oncológico.

1.2.1 *Microarrays* de ADN: porquê?

A expressão da informação genética contida na molécula de ADN ocorre em duas fases. Na primeira, a *transcrição*, uma parte da molécula de ADN é transcrita (copiada) em *ácido ribonucleico mensageiro* ou *mARN*, com base na propriedade de complementaridade. O mARN é uma molécula em tudo igual ao ADN, com as excepções de que é composta por uma única cadeia de nucleótidos e dos quais, a timina (T) é substituída pelo uracilo (U). Na segunda fase, a *tradução*, o mARN é descodificado para produzir uma proteína (sendo esta composta por uma sequência de aminoácidos). Por exemplo, uma sequência de CCC no mARN (e portanto GGG no ADN)[2] codifica o aminoácido *prolina*. Conhecer a abundância de transcrição dos genes (a partir da abundância de ARN mensageiro - mARN), tornou-se um ponto fundamental para responder às questões acima descritas. O processo de transcrição de uma sequência de ADN correspondente a um gene para uma sequência de mARN (que servirá como suporte para a produção de proteínas) é designado de expressão do gene (*gene expression*). Basicamente, o nível de expressão de um gene indica o número aproximado de cópias de mARN desse gene produzidas numa célula, que se pensa estar correlacionado com a quantidade de proteína correspondente produzida [5, 29]. A capacidade de monitorizar a expressão de um gene na fase da transcrição tornou-se possível graças à tecnologia dos *microarrays* de ADN. De facto, esta tecnologia, ao fornecer uma forma sistemática de monitorizar a variação de ADN e ARN, oferece a primeira grande esperança para uma visão global dos processos biológicos. Será certamente uma ferramenta essencial na investigação em biologia molecular e em diagnóstico clínico.

1.2.2 *Microarrays* de ADN: como?

Existem diferentes tipos de sistemas para obtenção de *microarrays*, entre os quais *microarrays* cADN (*cDNA* ou *complementary DNA*) e *microarrays* de oligonucleótidos (*Affymetrix, Inc.* em www.affymetrix.com), mas todos se baseiam na propriedade de complementaridade das moléculas de ADN. A descrição que se segue, centra-se na

[1] Não existe, actualmente, um método sistemático e único de diagnóstico; este requer a análise de diferentes especialistas em diferentes laboratórios altamente especializados.

[2] C de citosina e G de guanina

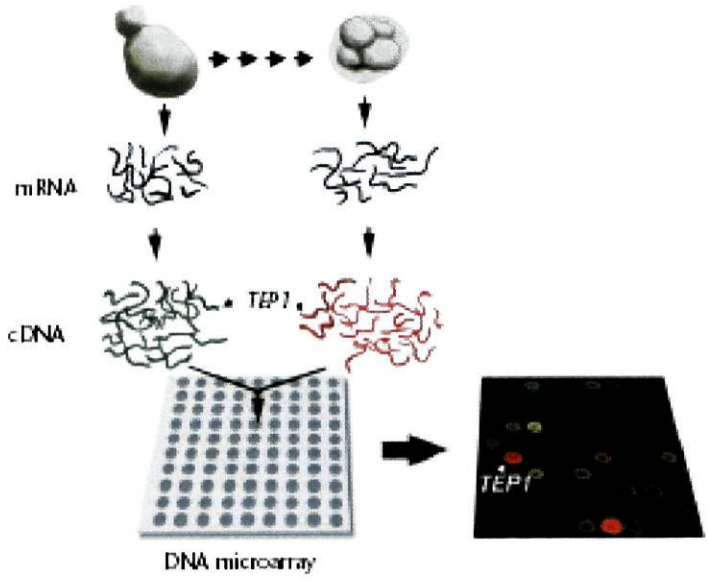


Figura 1.3: Esquema de um processo *microarray*, onde se pode observar a obtenção de mARN de células em duas condições distintas. Por transcrição inversa obtém-se o cADN, que é marcado com duas cores fluorescentes (vermelho/verde); este é misturado e hibridizado no *microarray* com as sequências de ADN em cada *spot* (cada *spot* contém várias sequências de um mesmo gene). A expressão de um determinado gene é medida a partir do quociente $Cy5/Cy3$ no respectivo *spot*. Figura adaptada de [1].

primeira metodologia (cADN).

Um *microarray* é essencialmente um suporte de vidro (como uma transparência) onde são colocadas em locais fixos e individuais (*spots*) várias porções de uma mesma sequência de ADN (por exemplo, correspondente a um determinado gene). Um só *microarray* pode conter milhares de *spots*. A forma mais popular de usar esta tecnologia é na comparação dos níveis de expressão dos genes em duas (ou mais) amostras diferentes, como por exemplo, em estado saudável e doente.

A abundância relativa de um dado gene nas duas condições pode ser determinada monitorizando a *hibridização* (ou complementarização) entre as sequências de ADN contidas no respectivo *spot* e as sequências de cADN (obtido por transcrição inversa de mARN das duas condições). Estas são etiquetadas com duas cores fluorescentes: vermelho ($Cy5$) para uma amostra (condição 1) e verde ($Cy3$) para a outra (condição 2). Após a aplicação do processo de hibridização de sequências complementares (veja-se a figura 1.3), cada *spot* é excitado por um laser. São então efectuadas medições das intensidades fluorescentes para cada cor. Se a quantidade de mRNA da amostra na condição 1 é abundante o *spot* será vermelho, se a quantidade na condição 2 for abundante o *spot* será verde, se em igual quantidade será amarelo e se em nenhum dos casos estiver presente será preto (veja-se a figura 1.4). É a partir das cores e das intensidades das fluorescências que, usando ferramentas de análise e processamento de imagem, é possível estimar os níveis de expressão dos genes: o quociente $Cy5/Cy3$ para

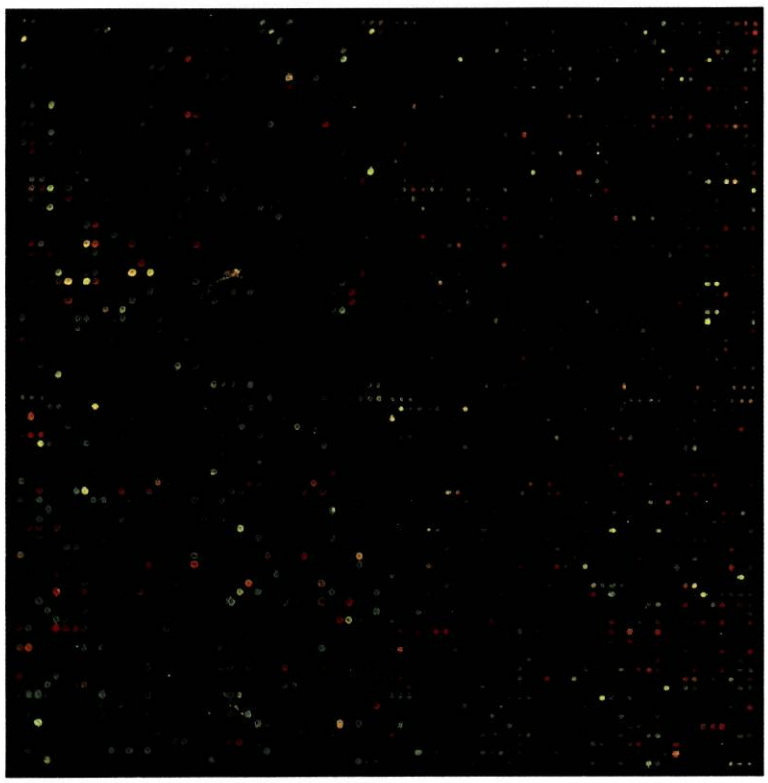


Figura 1.4: Imagem obtida pelo processo cADN, onde se podem observar as diferentes fluorescências. Esta imagem corresponde a todo o genoma da bactéria (levedura) do fermento(*Saccharomyces cerevisiae*).

cada *spot* é indicativo da abundância relativa da sequência de ADN correspondente nas duas amostras. Assim, se $Cy5/Cy3 > 1$ então o gene está sobreexpresso na condição 1, caso contrário está sobreexpresso na condição 2.
Uma descrição mais detalhada da tecnologia, dos diferentes processos e processamentos a efectuar até obter o resultado final - a matriz de expressões - pode ser vista em [1, 4, 5, 14, 23, 35]. Também está disponível na Internet uma animação da metodologia acima descrita em http://www.bio.davidson.edu/courses/genomics/chip/chip.html.

1.2.3 *Microarrays* de ADN: objectivos e aplicações.

O grande objectivo da investigação genética é, como já referido anteriormente, perceber como o genoma funciona. Agora que o genoma de vários seres eucariotas se encontra completamente determinado e que o humano para lá se encaminha, é essencial perceber e determinar as vantagens que poderão advir da grande quantidade de informação gerada. O padrão da expressão de um gene fornece informação sobre a sua função. De facto, é improvável que um gene expresso unicamente no rim esteja directamente envolvido na patologia da esquizofrenia. Os *microarrays* podem ser utilizados para monitorizar padrões de expressão em milhares de genes em simultâneo, gerando pistas acerca das suas funções e ajudando a identificar os alvos apropriados para a inter-

venção terapêutica. Desvios da fisiologia normal são frequentemente acompanhados por um conjunto de mudanças a nível histológico e bioquímico, incluindo mudanças nos padrões de expressão dos genes [1]. A aplicação mais atractiva desta tecnologia centra-se precisamente no estudo da expressão diferenciada dos genes em presença de doença; outro objectivo poderá ser a detecção de grupos de genes com padrões de expressão semelhantes. Se a sobreexpressão de certos genes está correlacionada com um certo cancro, poder-se-á explorar que outras condições afectam a expressão desses genes ou descobrir outros genes com perfis de comportamento semelhante. Poder-se-á, também, investigar que potenciais fármacos baixam os níveis de expressão desses genes, ou mesmo, em pós-análise, monitorizar as mudanças nos níveis de expressão em resposta a um determinado tratamento. A classificação de tumores cancerígenos é também uma das aplicações possíveis dos *microarrays*. O objectivo é criar uma estratégia geral e sistemática que permita melhorar e facilitar o diagnóstico oncológico, por forma a aplicar regimes terapêuticos correctos (quimioterapia por exemplo) a cada caso, maximizando a eficácia e minimizando a toxicidade.

O número e variedade de aplicações da tecnologia dos *microarrays* de ADN é ilimitada. A construção de uma base de dados com toda a informação gerada por esta tecnologia, ajudará a perceber a regulação dos genes, os processos metabólicos, os mecanismos genéticos das doenças e a resposta a tratamentos. A *Stanford Microarray Database*[3] (SMD) é um exemplo de uma base de dados deste tipo [28].

1.3 Metodologias e problemas

A tecnologia *microarray* gera matrizes de dados de grande dimensão. Habitualmente, matrizes com milhares (4 a 7) de linhas (representando os genes) e algumas dezenas (4 a 10) de colunas (representando as amostras) são obtidas[4] pelos processos já descritos. A ij-ésima entrada da matriz representa o nível de expressão do gene i na amostra j. No presente estudo (assim como na maior parte da literatura) consideram-se matrizes em que as amostras correspondem a tumores cancerígenos provenientes de diferentes pacientes. Mas então, que metodologias podem ser aplicadas para responder (ou pelo menos abrir caminho) às questões levantadas nas secções anteriores?
A análise deste tipo de dados pode ser efectuada de duas formas diferentes. Podemos considerar os genes como as nossas variáveis e os tumores como as observações e, por exemplo, estudar a semelhança entre as doenças e/ou construir predictores para novas amostras. A outra forma é considerar os genes como as observações e tentar, por exemplo, determinar que grupos de genes apresentam um padrão de expressão semelhante.
Em geral, os métodos podem ser divididos em dois grandes grupos: métodos de *análise discriminante* e métodos de *análise classificatória* [5, 26]. Os primeiros requerem,

[3] http://genome-www5.stanford.edu/MicroArray/MDEV/
[4] Num futuro não muito longínquo, estes valores serão ainda muito maiores [1]

para além da matriz de expressões, alguma informação adicional (exterior ao processo *microarray*) acerca dos genes ou das amostras que seja indicadora de uma classificação pré-existente. Normalmente, esta análise tem a construção de predictores como objectivo, como por exemplo, para novas amostras cujo diagnóstico não seja ainda conhecido (referido como *class prediction* em [13]). A regressão logística, redes neuronais, análise discriminante linear (LDA) ou quadrática (QDA), árvores de decisão, o classificador dos k-vizinhos mais próximos entre outros, são exemplos de métodos discriminantes. Por seu lado, os métodos de análise classificatória (*clustering*) não requerem qualquer tipo de informação adicional. O seu objectivo é encontrar padrões nos dados, tentando agrupar os objectos com características semelhantes (*class discovery* em [13]). Destacam-se os métodos de classificação não hierárquica como o k-médias e suas variantes, àrvores de classificação hierárquica, *self-organizing maps* (SOM) entre outros.

As dimensões das matrizes de dados em estudo colocam sérios problemas no que concerne à aplicação dos diferentes métodos. Consideremos, como exemplo, uma matriz de expressões com 6000 genes e 70 amostras (dois ou mais tipos de tumores). Vejamos então, que tipo de problemas podem ser colocados. Centrados no problema de classificação de tumores, consideremos os genes como as nossas variáveis. A elevada dimensionalidade (recorde-se que estamos num espaço 6000-dimensional) pode causar um fenómeno designado de *maldição da dimensão* (*curse of dimensionality*) [3]. Um exemplo da sua manifestação é o seguinte. Considere-se uma amostra de N pontos distribuídos uniformemente numa hiperesfera p-dimensional centrada na origem. A distância mediana da origem ao ponto da amostra mais próximo é dada pela expressão

$$d(p, N) = \left(1 - \frac{1}{2}^{1/N}\right)^{1/p}$$

No caso que estamos a considerar ($p = 6000$ e $N = 70$) $d(p, N) \approx .999$, ou seja, a maior parte dos pontos estão junto à superfície da esfera, o que afecta grandemente a sua distribuição. Desta forma vemos que a noção de distância e de vizinho mais próximo fica "distorcida". Assim, se neste exemplo quiséssemos usar a regra do vizinho mais próximo, esse vizinho encontra-se bastante afastado do ponto a prever; não faz portanto sentido o termo próximo.
Um outro exemplo é que a densidade de uma amostra é proporcional a $N^{1/p}$ em que p é a dimensão do espaço de variáveis e N o tamanho da amostra. Assim, se $N_1 = 100$ representa uma amostra densa num espaço unidimensional, então $N_{10} = 100^{10}$ é o tamanho necessário que uma amostra deve ter para obter a mesma densidade num espaço 10-dimensional (veja-se [15] para estes e outros exemplos).
Também a aplicabilidade de métodos como LDA ou QDA é reduzida quando o número de variáveis supera o número de observações. Assim, a construção de predictores fiáveis para o diagnóstico de tumores terá de seguir algum tipo de redução da dimensão, seja por escolha apropriada de genes mais predictivos, seja por combinações lineares das variáveis ou outros.

Este trabalho pretende, de alguma forma, fazer um estudo de diferentes formas de redução da dimensão num problema de *microarrays* de ADN, centrando-se no problema da construção de predictores para os tumores cancerígenos.

1.4 Ambiente computacional: o projecto R.

A versão inicial da linguagem R (www.r-project.org) foi desenvolvida por Ross Ihaka e Robert Gentleman da Universidade de Auckland. É uma linguagem semelhante ao S-Plus, ao ponto de se poder considerar como a versão grátis do S-Plus. Actualmente, o desenvolvimento do programa é dirigido por uma dúzia de pessoas provenientes de várias instituições de diferentes países. O espírito do R é do designado *open source system*, isto é, todo o código encontra-se ao alcance de cada utilizador; esta é, de facto, uma das suas grandes vantagens pois permite a identificação de *bugs* ou inadequações, levando a um melhoramento sistemático das funções.
O R pode ser obtido a partir do CRAN (Comprehensive R Archive Network) em http://cran.r-project.org bem como um vasto conjunto de bibliotecas desenvolvidas pelos seus utilizadores.
Para o caso particular dos *microarrays* de ADN e devido à grande dimensão destas estruturas, é desenvolvido paralelamente ao R um projecto designado Bioconductor, www.bioconductor.org. Um dos objectivos deste projecto consiste no desenvolvimento de bibliotecas específicas para o manuseamento deste tipo de dados.
A literatura para a linguagem R é vasta e variada. Cada biblioteca vem acompanhada de documentos explicativos e com exemplos. O próprio R contém alguns manuais, destacando-se o introdutório à linguagem [34]. A partir do CRAN podem também ser obtidas várias publicações, a *newsletter* do R e pequenos manuais desenvolvidos por utilizadores [19]. A biblioteca MASS foi baseada no livro de Ripley *et al.* [27].
Todo o trabalho prático e computacional desenvolvido ao longo deste trabalho foi produzido com as ferramentas da linguagem R. As funções programadas de maior importância encontram-se em Anexo.

Capítulo 2

Selecção de genes predictivos

A selecção de genes predictivos assenta na ideia de que boas predições baseiam-se em bons predictores. Como já foi referido no capítulo anterior, desvios da fisiologia normal (como por exemplo devidos a cancro) provocam alterações a vários níveis, entre os quais, alterações nos padrões de expressão dos genes. No entanto, apesar do elevado número de genes monitorizados, apenas alguns exibem um comportamento diferenciado para cada classe; aliás, um grande número exibe padrões de expressão praticamente constantes ao longo das diferentes classes. Estes genes não apresentam um poder discriminativo pelo que podem ser extraídos. Surge então a necessidade de seleccionar os genes (variáveis) mais informativos (predictivos) da distinção entre classes.
Mas então que padrões de expressão devemos procurar e reter? A resposta é, de certa forma, intuitiva. Por um lado, a expressão de um gene predictivo deve ser diferente de classe para classe, isto é, um gene deve ser sobreexpresso numa classe e subexpresso na outra (no caso de duas classes). Por outro lado, dentro de cada classe a variação deve ser a menor possível, ou seja, não deve haver grandes variações em torno da média. Note-se que se tal variação existisse, não estaríamos a distinguir entre classes, mas quando muito entre pacientes.
Sucintamente, a selecção deve reter os genes com uma grande gama de valores, mas que esta variação se deva essencialmente à diferença entre classes.

As secções seguintes exploram algumas das possibilidades existentes para a extracção de genes predictivos. Na primeira secção expõe-se uma forma de pré-processamento que permite filtrar uma grande quantidade de genes. Na segunda e terceira secções exploram-se métodos baseados em índices estatísticos para a selecção de genes predictivos; a primeira destas secções dedica-se ao caso de duas classes e as seguintes ao caso multiclasses.

2.1 Filtragem inicial

No conjunto de dados LEUCEMIA[1] é efectuado em várias referências uma pré-filtragem de genes (de acordo com Golub *et al.* [13]). Esta filtragem inicial tem o objectivo de eliminar da análise genes com pouca (se alguma) variação na expressão ao longo das amostras. Embora não sendo uma forma completamente correcta de seleccionar genes (recorde-se, variáveis), pode ser considerado como um passo inicial. Aliás, a filtragem é sempre uma boa ideia, evitando assim o uso de uma grande "massa"de genes com pouca variação, podendo afectar as análises.
O procedimento para esta pré-filtragem é o seguinte. Aplicam-se limites (*thresholds*) máximo de 16000 e mínimo de 100 para os níveis de expressão da matriz. Os genes cujos níveis max/min e $max - min$ são menores que 5 e 500 respectivamente, são posteriormente rejeitados. Este procedimento permite reduzir o número de genes em estudo de 7129 para 3571. Golub *et al.* [13] justificam o limite mínimo positivo com o facto de que valores negativos de expressão seriam difíceis de interpretar[2], no entanto não encontrei nas diversas referências e contactos que estabeleci nenhuma justificação para os restantes valores acima descritos.

Este tipo de pré-processamento não é exclusivo dos dados LEUCEMIA. Com efeito, em grande parte dos conjuntos de dados existentes na literatura é efectuado este género de filtragem como passo inicial, que permite excluir da análise genes com pouca variação. No entanto, os dados já são fornecidos após este procedimento. Note-se que acaba por ser um procedimento extremamente importante, pois a redução efectuada é, de certa forma, drástica (para o conjunto LEUCEMIA obteve-se uma redução para cerca de metade dos genes inicialmente monitorizados).

[1] Descrito mais adiante.
[2] De acordo com a definição de expressão de um gene.

2.2 Selecção de genes predictivos para o caso de duas classes

Para um problema de duas classes, é necessário construir um critério que permita ordenar os genes de acordo com as ideias expostas no início do capítulo e que podem ser ilustradas na figura 2.1. A escolha dos genes centra-se precisamente naqueles com um comportamento o mais semelhante ao de baixo.

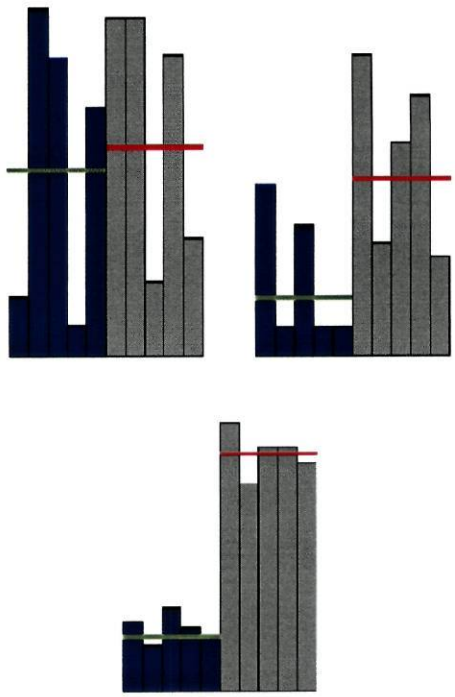


Figura 2.1: Perfis exemplificativos de 3 genes em 10 amostras (5 da classe azul e 5 da classe cinzenta). As linhas horizontais representam as respectivas médias em cada classe. Esta figura mostra a ideia da selecção de genes predictivos. O gene representado em cima à esquerda não será um bom predictor, pois as médias para as duas classes são muito próximas. Os restantes, possuem médias para cada classe bem separadas. No entanto, o gene em baixo será um melhor predictor, pois tem uma menor variação em torno de cada média.

2.2.1 Métrica de correlação e estatística t^*

Sejam $\hat{\mu}_1$ ($\hat{\mu}_2$ resp.) e $\hat{\sigma}_1$ ($\hat{\sigma}_2$ resp.) a média e desvio padrão respectivamente de um gene[3] g para as amostras de classe 1 (classe 2 resp.). Golub *et al.* [13] propõem[4] o índice

$$P(g,c) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_1 + \hat{\sigma}_2} \tag{2.1}$$

[3]Na verdade, o logaritmo dos níveis de expressão.

[4]Os autores experimentaram também a distância euclideana, de Battacharyya e Manhattan, bem como o coeficiente de correlação de Pearson. $P(g,c)$ obteve a melhor performance.

que designaram de *métrica de correlação*[5] e que mede a separação relativa entre classes [13, 29].
Este índice permite medir o grau de semelhança entre o nível de expressão de um gene e um vector **c** de expressão idealizada que possui para uma classe níveis uniformemente altos e para outra uniformemente baixos (ver figura 2.2).

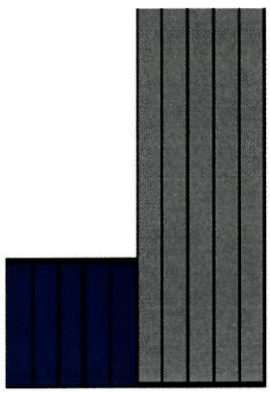


Figura 2.2: Perfil ideal de expressão. Um gene predictivo deverá ter um comportamento o mais próximo deste: níveis de expressão uniformemente altos numa classe e uniformemente baixos noutra.

Desta forma, os valores máximos de $P(g,c)$ correspondem a genes sobreexpressos na classe 1 e os valores máximos de $-P(g,c)$ correspondem a genes sobreexpressos na classe 2.
Antes de escolher os genes predictivos, os autores, baseados na ideia de que nem todas as distinções são determinadas unicamente pela expressão genética, investigam se de facto existem genes passíveis de serem bons predictores. Para tal aplicam uma metodologia, que designam de *análise de vizinhança*, e que consiste em contar o número de genes numa vizinhança fixa de um vector de expressão idealizada **c** (ver figura 2.2). Comparando com o número de genes contidos numa vizinhança do mesmo tamanho em torno de várias permutações aleatórias de **c**, podem concluir se a vizinhança de **c** contém mais genes do que os esperados. Se sim, podem concluir que a distinção de classes representada por **c** pode ser predicta a partir da expressão genética [13, 29]. Para o caso particular dos dados que estudaram, LEUCEMIA, Golub *et al.* [13] determinaram que cerca de 1100 genes eram muito correlacionados com a distinção pretendida (ALL *versus* AML). Para construir o seu predictor, os autores escolhem p^* genes mais predictivos que correspondem a $p^*/2$ mais sobreexpressos na classe 1 e $p^*/2$ mais sobreexpressos na classe 2.
O problema prende-se agora com a escolha de p^*. Terá de existir um compromisso entre a quantidade de informação e robustez ganha ao seleccionar mais genes e a quantidade de ruído adicionado. Golub *et al.* [13] variaram p^* entre 10 e 200 e verificaram que o seu modelo predictivo não era muito sensível ao número exacto de genes escolhidos. A opção recaiu por $p^* = 50$.

[5] Ao contrário dos coeficientes de correlação habituais, $P(g,c)$ pode tomar valores fora de $[-1,1]$

Por seu lado, Nguyen *et al.* [25] baseiam a sua selecção de genes informativos no que designam[6] de estatística t

$$t^* = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{\frac{\hat{\sigma}_1^2}{N_1} + \frac{\hat{\sigma}_2^2}{N_2}}} \tag{2.2}$$

com N_k, $\hat{\mu}_k$ e $\hat{\sigma}_k^2$ o tamanho, média e variância da classe $k = 1, 2$. O esquema é igual ao anterior. Calcular t^* para cada gene e reter os $p^*/2$ genes com os maiores valores e os $p^*/2$ genes com os menores valores, o que corresponde, como anteriormente, aos $p^*/2$ genes sobreexpressos na classe 1 e $p^*/2$ genes sobreexpressos na classe 2 respectivamente. O valor p^* base foi também de 50 genes, embora nas experiências conduzidas com os seus predictores tenham efectuado comparações para $p^* = 50, 100, 500, 1000, 1500$.

Note-se que, de facto, tanto para o caso da métrica de correlação (2.1) como para a estatística t^* (2.2), estamos a seleccionar genes de acordo com os critérios pretendidos. Escolher os valores máximos de $P(g, c)$ ou t^* corresponde a seleccionar os genes com $\hat{\mu}_1 > \hat{\mu}_2$ e $\hat{\sigma}_1$ e $\hat{\sigma}_2$ pequenos, ou seja, sobreexpressos na classe 1 e com pouca variação dentro de cada classe (ou a menor possível). Por outro lado, escolher os valores mínimos de $P(g, c)$ (máximos de $-P(g, c)$) ou t^* corresponde a seleccionar genes sobreexpressos na classe 2, pois $\hat{\mu}_2 > \hat{\mu}_1$ e $\hat{\sigma}_1$ e $\hat{\sigma}_2$ também pequenos.

A figura 2.3 mostra o chamado *heat map* para os níveis de expressão dos 50 genes predictivos seleccionados por Golub *et al.* [13]. Como se pode observar os primeiros 25 estão sobreexpressos na classe 1 (ALL) e subexpressos na classe 2 (AML). Os segundos 25 estão sobreexpressos na classe 2 e subexpressos na classe 1.

2.2.2 Discussão e sugestões

O problema da selecção de genes predictivos pode ser formulado como um teste estatístico. Suponhamos que a distribuição de cada classe e para cada gene é Normal ou aproximadamente Normal com variâncias iguais[7]. Então, um gene será tanto mais predictivo quanto mais rejeita a hipótese nula[8] do teste

$$\begin{aligned} H_0 &: \mu_1 - \mu_2 = 0 \\ H_1 &: \mu_1 - \mu_2 \neq 0 \end{aligned} \tag{2.3}$$

em que μ_1 e μ_2 são as médias populacionais para a classe 1 e 2 respectivamente. A solução não é mais do que um teste t em que rejeitamos os genes com $|t| < t_{\alpha/2, N_1+N_2-2}$. Rigorosamente, devido ao elevado número de testes a realizar, o nível de significância

[6] A forma como os autores descrevem o seu índice parece indicar que estarão a considerar que t tem distribuição t-Student, mas tal não é verdade. Veja-se a discussão mais adiante.

[7] É razoável assumir a igualdade das variâncias. Veja-se a figura 2.2.

[8] Quanto mais significativa é a diferença de médias.

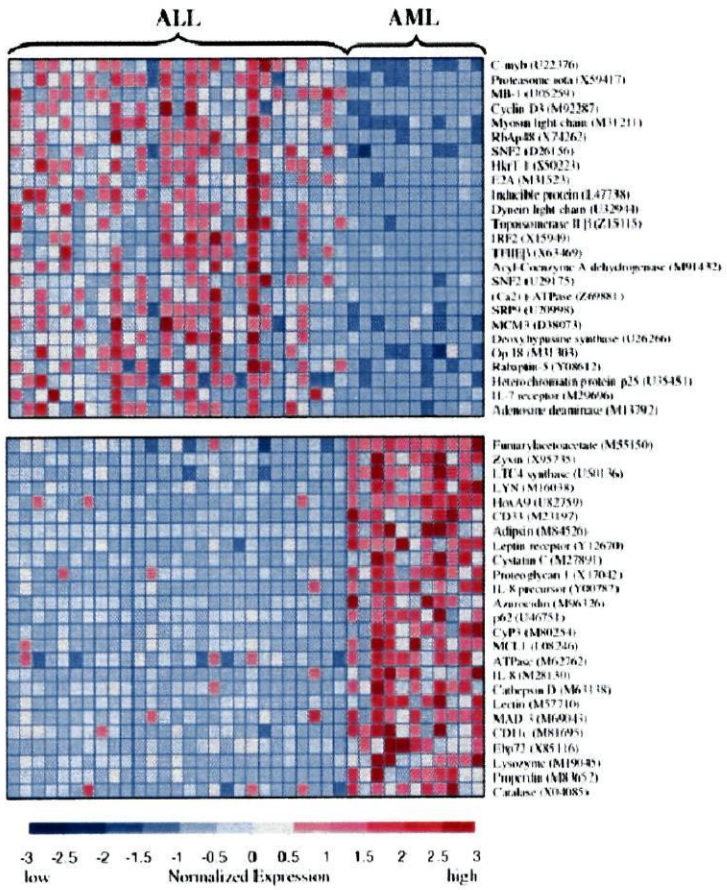


Figura 2.3: Genes discriminativos das duas classes ALL e AML. A figura contém os níveis de expressão normalizados para os 50 genes predictivos de Golub *et al.* [13]. Níveis maiores que a média estáo coloridos em tons de vermelho e aqueles menores que a média em tons de azul. A escala indica os desvios padrões abaixo ou acima da média. Os primeiros 25 genes apresentam uma sobreexpressão na classe ALL, enquanto que os segundos estão sobreexpressos na classe AML. Figura adaptada de [13].

a utilizar em cada teste teria de ser diminuído por forma a termos um nível final de α. Este pode ser um critério de escolha do número p^* de genes predictivos. Recorde-se que para os casos anteriores (Golub *et al.* [13] e Nguyen *et al.* [25]) a escolha era de certa forma arbitrária. A estatística para o teste (2.3) é, sob H_0 [9, 22]

$$t = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\sqrt{S_p^2 \left(\frac{1}{N_1} + \frac{1}{N_2}\right)}} \sim t_{N_1+N_2-2} \tag{2.4}$$

com

$$S_p^2 = \frac{(N_1 - 1)\hat{\sigma}_1^2 + (N_2 - 1)\hat{\sigma}_2^2}{(N_1 - 1) + (N_2 - 1)}.$$

As estatísticas (2.1) e (2.2) propostas, parecem, de algum modo, aplicar a metodologia de um teste t. No entanto, existem alguns pormenores que interessa ressalvar. Relativamente à segunda, a estatística t^* de Nguyen *et al.* [25], a sua distribuição não é, sob

H_0, $t-Student$. De facto, a distribuição em causa depende de N_1,N_2 e σ_1^2/σ_2^2, o que implica que a probabilidade de rejeição varia com σ_1^2/σ_2^2 [9]. A distribuição de (2.2) é designada **distribuição de Behrens-Fisher** e o problema (2.3) tem, neste caso, soluções muito particulares (veja-se [9]).
Quanto à métrica de correlação (2.1), Dudoit *et al.* [10] fazem uma pequena observação. Considere-se o discriminante de máxima verosimilhança [10]. Este, atribui a uma observação $\mathbf{x} = (x_1, \ldots, x_p)$ a classe k tal que $pr(\mathbf{x}|k) = max_j pr(\mathbf{x}|j)$ em que $pr(\mathbf{x}|j)$ é a densidade condicional da classe j. Para o caso especial em que $\mathbf{x}|k$ tem distribuição $N(\mu_k, \Sigma)$, $\Sigma = diag\left(\sigma_1^2, \ldots, \sigma_p^2\right)$ e $k = 2$ a regra classifica $\mathbf{x}$ na classe 1 se

$$\sum_{j=1}^{p} \frac{(x_j - \bar{x}_{2j})^2}{\hat{\sigma}_j^2} \geq \sum_{j=1}^{p} \frac{(x_j - \bar{x}_{1j})^2}{\hat{\sigma}_j^2}$$

ou seja, se

$$\sum_{j=1}^{p} \frac{(\bar{x}_{1j} - \bar{x}_{2j})}{\hat{\sigma}_j^2} \left(x_j - \frac{(\bar{x}_{1j} + \bar{x}_{2j})}{2}\right) \geq 0 \tag{2.5}$$

A regra anterior é muito semelhante ao esquema de votação pesada[9] de Golub *et al.* [13] com a excepção de que o primeiro quociente é substituído por

$$\frac{\bar{x}_{1j} - \bar{x}_{2j}}{\hat{\sigma}_{1j} + \hat{\sigma}_{2j}}$$

que corresponde a $P(g_j, c)$. Duas observações são feitas por Dudoit *et al.* [10]. Por um lado, $\hat{\sigma}_{1j} + \hat{\sigma}_{2j}$ é uma forma pouco usual de determinar o desvio padrão de uma diferença; por outro, o uso de desvios no lugar de variâncias em (2.5) produz as unidades erradas.

No início desta discussão, supôs-se que a distribuição de cada classe era Normal ou aproximadamente Normal. Mas será razoável assumir uma condição tão forte para os nossos dados? Até que ponto se poderá apoiar no Teorema do Limite Central (TLC)? De facto, o TLC tem pouca força nos estudos *microarray*, porque a dimensão das amostras é pequena. Por outro lado, este tipo de dados é tipicamente assimétrico [23, 25], o que viola a condição de Normalidade na maior parte dos testes estatísticos paramétricos. Para contornar este problema[10] é usual aplicar uma transformação logarítmica aos dados para que a sua distribuição seja mais próxima da Normal.

2.3 Selecção de genes predictivos para mais do que 2 classes

As estatísticas exploradas na secção anterior são direccionadas para o problema de classificação binário (2 classes). No entanto, na prática, podemos ter situações em

[9] Apresentado mais adiante.
[10] Para além de outras particularidades deste tipo de dados (veja-se [23]).

que o problema possui várias classes[11]. O problema para mais do que duas classes pode ser visto como uma extensão dos conceitos utilizados anteriormente: continuamos a pretender genes com expressão diferenciada entre classes. No entanto, novas estatísticas têm de ser utilizadas. Suponhamos nesta secção que dispomos de K classes.

2.3.1 *Um contra todos*

Uma das opções mais intuitivas para resolver o problema multiclasses é considerar o método *um contra todos* (*one against all*) que consiste basicamente em aplicar as metodologias para duas classes, considerando como classe 1 uma das K classes e como classe 2 todas as outras. Em K passos podemos seleccionar os genes mais predictivos de cada classe. Por exemplo, suponhamos as classes A,B e C. Considera-se A como a classe 1 e $B \cup C$ como a classe 2. Aplica-se a métrica de correlação e seleccionam-se os genes mais predictivos de A. No próximo passo, considera-se B como a classe 1 e $A \cup C$ como a classe 2 e determinam-se os genes mais predictivos para B. Analogamente para C. Este método permite determinar de uma forma eficaz (pelo menos tão eficaz quanto os casos de duas classes) os genes mais predictivos para o conjunto de dados em estudo. No entanto, é um método muito mais exigente computacionalmente do que os seguintes, principalmente se o número de classes for elevado.

2.3.2 Análise de variância. Estatística F

O problema da selecção de genes para o caso de múltiplas classes pode ser formulado da seguinte forma. Para um dado gene, pretendemos testar a hipótese

$$\begin{aligned} H_0 &: \mu_1 = \mu_2 = \ldots = \mu_K \\ H_1 &: \text{pelo menos uma média difere} \end{aligned} \tag{2.6}$$

O objectivo é excluir os genes para os quais a hipótese nula não pode ser rejeitada, ou seja, genes que não possuem poder discriminativo entre classes. Seja N_i o número de elementos da classe i, K o número de classes e suponhamos que os níveis de expressão nas diferentes classes são Normalmente distribuídos e que as observações são independentes e $N(\mu_i, \sigma^2)$. Então o problema (2.6) pode ser resolvido com o método[12] de **análise de variância** (anova) [9, 22]. A hipótese nula em (2.6) é rejeitada quando

$$\frac{MQE}{MQD} > F_\alpha \tag{2.7}$$

[11] Vários tipos de tumor ou subtipos de um mesmo tumor que interessa identificar.
[12] Note-se que σ^2 é igual para todas as classes.

em que $F_\alpha = F_{K-1,\sum N_i - K, 1-\alpha}$ e

$$\begin{aligned} MQE &= \frac{SQE}{K-1} = \frac{\sum_{i=1}^{K} N_i \left(\overline{X}_i - \overline{\overline{X}}\right)^2}{K-1} \\ MQD &= \frac{SQD}{\sum N_i - K} = \frac{\sum_{i=1}^{K}\sum_{j=1}^{N_i} \left(X_{ij} - \overline{X}_i\right)^2}{\sum N_i - K} \\ \overline{X}_i &= \frac{1}{N_i}\sum_{j=1}^{N_i} X_{ij} \\ \overline{\overline{X}} &= \frac{\sum_{i=1}^{K} N_i \overline{X}_i}{\sum_{i=1}^{K} N_i} \end{aligned}$$

Assim, seleccionamos os genes que, para um dado nível[13] α, rejeitam a hipótese nula, ou seja, têm $MQE/MQD > F_\alpha$. Se tal não for possível, em virtude de todos os valores da estatística serem inferiores ao quantil $1-\alpha$ da distribuição (o que não se espera à partida), podemos optar por escolher, à luz de metodologias anteriores, os p^* genes com maior valor de MQE/MQD.

2.3.3 Rácio BSS/WSS

Dudoit *et al.* [10] propõem para a selecção de genes predictivos o rácio *soma dos quadrados entre classes* sobre *soma dos quadrados dentro das classes*

$$\frac{BSS(j)}{WSS(j)} = \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \tag{2.8}$$

onde $\bar{x}_{kj}$ representa a expressão média do gene j nas amostras da classe k e $\bar{x}_{.j}$ a expressão média do gene j em todas as amostras. Ao escolher os p^* genes com maior rácio BSS/WSS, estamos a escolher as variáveis que permitem uma maior separação inter-classes e uma menor separação intra-classes ("cf. análise discriminante linear de Fisher, [8]").

Seja K o número de classes. Manipulando um pouco a expressão (2.8) temos

$$\begin{aligned} \frac{BSS(j)}{WSS(j)} &= \frac{\sum_i \sum_k I(y_i = k)(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_i \sum_k I(y_i = k)(x_{ij} - \bar{x}_{kj})^2} \\ &= \frac{\sum_{k=1}^{K}\sum_{i=1}^{N_k}(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ij} - \bar{x}_{kj})^2} = \frac{\sum_{k=1}^{K} N_k(\bar{x}_{kj} - \bar{x}_{.j})^2}{\sum_{k=1}^{K}\sum_{i=1}^{N_k}(x_{ij} - \bar{x}_{kj})^2} \\ &= \frac{SQE}{SQD} = \frac{K-1}{\sum_k N_k - K}\frac{MQE}{MQD} \end{aligned}$$

[13] Em boa verdade, também aqui o nível de significância deverá ser ajustado para a multiplicidade de testes que são efectuados

Verifica-se que o rácio BSS/WSS é muito semelhante a uma estatística F. Mais, é a estatística F multiplicada por uma constante. Assim, esperam-se resultados semelhantes com os dois métodos.

2.4 Método do centróide encolhido mais próximo

O método proposto por Tibshirani *et al.* [32, 33] baseia-se no classificador do protótipo mais próximo. Este consiste em atribuir a uma nova amostra a classe cujo protótipo (neste caso, centro) lhe está mais próximo. Na verdade, o método proposto é um desenvolvimento deste último e usa centróides "encolhidos"como protótipos, resultantes da identificação dos subconjuntos de genes que melhor caracterizam cada classe[14]. Introduzam-se algumas definições para clarificar a exposição. Seja $\mathbf{X}$ uma matriz de expressões de dimensão $p \times n$ e considere-se

$$\begin{aligned} x_{ij} &\rightarrow \text{expressão do gene } i \text{ na amostra } j \\ \bar{x}_{ik} &\rightarrow \text{a } i\text{-ésima componente do centróide para a classe } k \\ \bar{x}_i &\rightarrow \text{a } i\text{-ésima componente do centróide total} \\ C_k &\rightarrow \text{índices das } n_k \text{ amostras da classe } k \\ K &\rightarrow \text{número de classes} \end{aligned}$$

Consideremos os índices

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \left(s_i + s_0\right)} \tag{2.9}$$

em que

$$s_i^2 = \frac{1}{n-K} \sum_k \sum_{j \in C_k} \left(x_{ij} - \bar{x}_{ik}\right)^2 \tag{2.10}$$

$$m_k = \sqrt{1/n_k + 1/n} \tag{2.11}$$

O valor de s_0 é incluído no denominador para precaver contra a possibilidade de surgirem grandes valores de d_{ik} devido a genes com níveis de expressão muito baixos. Este valor, igual para todas as classes, é a mediana dos valores s_i. Reescrevendo (2.9), temos

$$\bar{x}_{ik} = \bar{x}_i + m_k \left(s_i + s_0\right) d_{ik} \tag{2.12}$$

O método pretende "encolher"d_{ik} obtendo novos índices d'_{ik} tais que[15]

$$d'_{ik} = sign(d_{ik}) \left(|d_{ik}| - \Delta\right)_+ \tag{2.13}$$

em que Δ é uma quantidade (*threshold*) positiva. Se $d'_{ik} < 0$, então estabelece-se $d'_{ik} = 0$. Resultam, assim, novos centróides (encolhidos)

$$\bar{x}'_{ik} = \bar{x}_i + m_k \left(s_i + s_0\right) d'_{ik} \tag{2.14}$$

[14] Esta é, aliás, a ideia presente na maior parte dos artigos descritos.

[15] $+$ significa *parte positiva*, i.e, $t_+ = t$ se $t > 0$ ou $t_+ = 0$ se $t \leq 0$

Seja $\mathbf{x}^* = (x_1^*, \ldots, x_p^*)$ o vector de expressões para uma nova amostra. Esta é classificada na classe de cujo centróide encolhido está mais próxima, ou seja,

$$\mathbf{x}^* \text{ pertence à classe } l \text{ se } \delta_l(\mathbf{x}^*) = \min_k \delta_k(\mathbf{x}^*) \tag{2.15}$$

em que

$$\delta_k(\mathbf{x}^*) = \sum_{i=1}^{p} \frac{\left(x_i^* - \bar{x}_{ik}'\right)^2}{\left(s_i + s_0\right)^2} - 2\log \pi_k \tag{2.16}$$

é o discriminante para a classe k. O primeiro termo em (2.16) não é mais do que o quadrado da distância estandardizada da nova amostra $\mathbf{x}^*$ ao k-ésimo centróide encolhido.

Este discriminante é uma forma restrita do método de análise discriminante linear. Com efeito, se assumirmos que $\mathbf{x}|k$ tem distribuição $N_p(\mu_k, \Sigma)$, se estimarmos μ_k por $(\bar{x}_{1k}', \ldots, \bar{x}_{pk}')$ e assumirmos que Σ é diagonal com estimativa

$$\hat{\Sigma} = \begin{pmatrix} (s_1 + s_0)^2 & & \\ & \ddots & \\ & & (s_p + s_0)^2 \end{pmatrix}$$

o que obtemos é precisamente a expressão (2.16). Também é possível construir estimativas das probabilidades condicionais *a posteriori*, usando os discriminantes

$$\hat{p}_k(\mathbf{x}^*) = \frac{e^{-\frac{1}{2}\delta_k(\mathbf{x}^*)}}{\sum_l e^{-\frac{1}{2}\delta_l(\mathbf{x}^*)}}$$

A grande vantagem deste método é que relativamente à classificação de novas amostras, grande parte dos genes são eliminados. Porquê? Repare-se que, se para um dado gene i, Δ é suficientemente grande para que $d_{ik}' = 0\ \forall k$, então por (2.14) o seu centróide para cada classe é igual a $\bar{x}_i$. Vejamos com um exemplo prático.

Sem perda de generalidade, suponhamos $K = 2$ e que o gene 1 tem $d_{11}' = d_{12}' = 0$. Então $\bar{x}_{1k}' = \bar{x}_1$ para $k = 1, 2$ e

$$\delta_1(x^*) = \frac{\left(x_1^* - \bar{x}_1\right)^2}{\left(s_1 + s_0\right)^2} + \sum_{i=2}^{p} \frac{\left(x_i^* - \bar{x}_{i1}'\right)^2}{\left(s_i + s_0\right)^2} - 2\log \pi_1$$

$$\delta_2(x^*) = \frac{\left(x_1^* - \bar{x}_1\right)^2}{\left(s_1 + s_0\right)^2} + \sum_{i=2}^{p} \frac{\left(x_i^* - \bar{x}_{i2}'\right)^2}{\left(s_i + s_0\right)^2} - 2\log \pi_2$$

Daqui vemos que de acordo com a regra (2.15) o gene 1 não contribui para a classificação de x^*, e, portanto, é desprezável.

Para determinar o Δ óptimo, efectua-se validação cruzada sobre o conjunto de treino ou recorre-se a um conjunto de validação independente.

Note-se que d_{ik} é semelhante a uma estatística t (um pouco modificada devido a s_0) para o teste

$$\begin{aligned} H_0 &: \mu_{ik} - \mu_i = 0 \\ H_1 &: \mu_{ik} - \mu_i \neq 0 \end{aligned}$$

ou seja, para a comparação da classe k com o centro total μ para cada gene i. Na verdade, o método não está mais do que a escolher genes com os maiores valores da estatística t, mas com critério de paragem o erro de classificação.

Os resultados obtidos com este método são satisfatórios. Por exemplo, dos 4026 genes que compunham o conjunto de dados LINFOMA, foi possível reduzir para 2938 activos[16] com $\Delta = 0.918$ (veja-se a figura 2.4). Embora esta redução se possa considerar significativa, o número de genes ainda activos não é competitivo (quando comparado com outros métodos) tanto a nível de manipulação como de interpretação biológica. Acontece que muitos dos genes que permaneceram activos poderão não ser necessários para uma classificação eficaz. Isto deve-se ao facto de que algumas classes poderão estar mais afastadas do centro total do que outras e, portanto, são mais fáceis de identificar. Surge então a ideia de um método adaptativo que varie o *threshold* Δ de forma diferente para cada classe, por forma a minimizar o número total de genes activos necessários para alcançar uma determinada taxa de erro. Este método consiste, basicamente, numa pequena alteração no cálculo das diferenças d_{ik}.
Consideremos o vector $(\theta_1, \ldots, \theta_K)$ que contém os coeficientes de escala para cada classe. Inicialmente, faça-se $\theta_k = 1 \; \forall k$ e

$$d_{ik} = \frac{\bar{x}_{ik} - \bar{x}_i}{m_k \theta_k \left(s_i + s_0\right)} \tag{2.17}$$

Aplica-se então, o seguinte procedimento [32]

1. Correr o algoritmo do centróide encolhido.
2. Encontrar a classe k com a maior média de erros no treino. Esta média é efectuada sobre todos os valores Δ utilizados.
3. Diminuir θ_k em 10% e reescalar todos os θ_j de tal forma que $min_j \; \theta_j = 1$;
4. Repetir os passos anteriores para m iterações e determinar a solução que obtém o menor erro médio entre todos os valores de $(\theta_1, \ldots, \theta_K)$ visitados.

Dos 4026 genes que compunham o conjunto de dados LINFOMA, apenas 48 se mantiveram activos sem aumentar o erro no conjunto de teste (veja-se a figura 2.4). Os valores dos parâmetros foram $\Delta = 4.41$ e $(\theta_{DLBCL}, \theta_{FL}, \theta_{B-CLL}) = (1.88, 1.00, 1.52)$. Daqui vemos que a classe mais problemática (ou seja, onde se cometem mais erros) é a segunda, pois foram aplicadas maiores penalizações às outras duas classes.

[16] Com pelo menos um $d'_{ik} \neq 0$

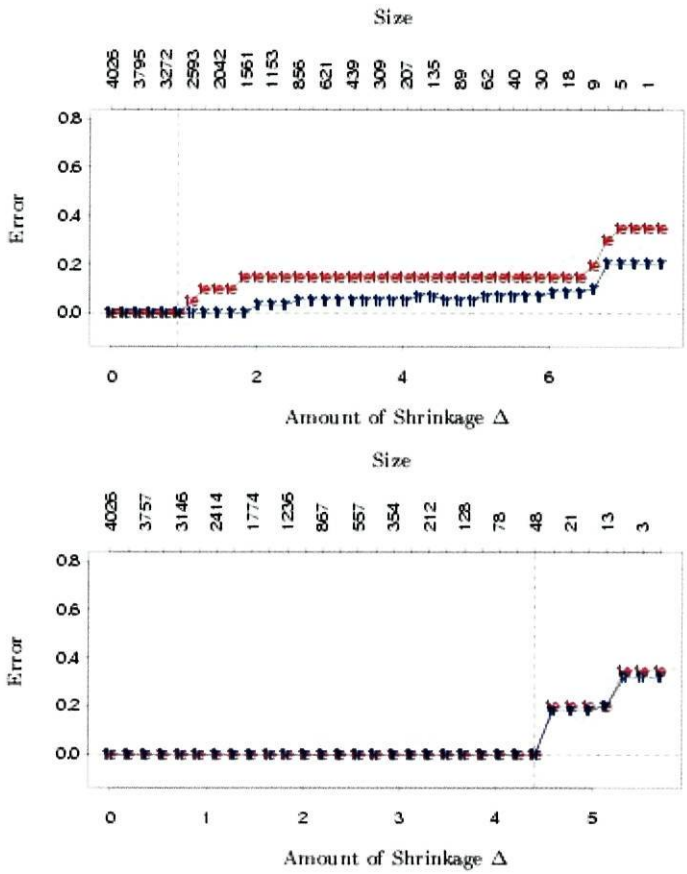


Figura 2.4: Erro de treino (tr) e erro de validação cruzada (te) com a variação do parâmetro Δ. Em cima, o método não adaptativo obtém uma solução com $\Delta = 0.918$ e 2938 genes activos. Em baixo, o método adaptativo obtém uma solução com $\Delta = 4.41$ e somente 48 genes activos, com a mesma taxa de erro que em cima. Figura adaptada de [32].

Capítulo 3

Combinações lineares

Os *microarrays* de ADN são actualmente empregues na criação de mecanismos de classificação de amostras em categorias, como por exemplo o tipo de tumor. Como já vimos, esta tecnologia gera informação em que o número de variáveis excede largamente o número de observações e, portanto, as metodologias habituais de classificação não funcionam. No capítulo anterior exploraram-se vários métodos de redução de variáveis baseando a escolha selectiva de genes na sua expressão diferenciada nas diferentes classes. Outras metodologias baseiam-se em projecções optimizadas em espaços de dimensão inferior. Aqui, as novas variáveis são habitualmente designadas de componentes. Estas são, na maior parte dos casos, combinações lineares das variáveis originais que explicam nalgum sentido a estrutura dos dados. A redução da dimensão é obtida com uma escolha conveniente do número de componentes a utilizar. Estes métodos podem também servir como complemento aos abordados no capítulo anterior. Veja-se por exemplo o caso de Golub *et al.* [13]. Foi construído um predictor (por votação pesada) utilizando os 50 genes predictivos pré-seleccionados pela métrica de correlação (2.1). No entanto, o uso de métodos tradicionais, como LDA, não seria aconselhado, pois o número de amostras no conjunto de treino (38) ainda é inferior ao número de variáveis seleccionadas. Aplicar um método de combinações lineares permitirá, por um lado, reduzir o número de variáveis e por outro, poderá revelar estruturas nos dados que não eram detectadas com as variáveis originais (note-se que as componentes permitem colocar em interação as variáveis originais). Este capítulo explora precisamente algumas das opções existentes: análise em componentes principais, mínimos quadrados parciais (componentes PLS) e variáveis canónicas.

3.1 Análise em componentes principais

A análise em componentes principais (também designada de transformação de Karhunen-Loève ou transformação de Hotelling) é amplamente usada em várias áreas, tais como processamento de sinal e estatística.
O objectivo da análise em componentes principais (ACP) é explicar, através de algumas combinações lineares, a estrutura de variância-covariância de um conjunto de variáveis. A ideia base de ACP é encontrar combinações lineares (componentes) não correlacionadas das variáveis originais $X_1, ..., X_p$ que explicam o máximo de variância. Geometricamente, estas combinações lineares correspondem à construção de um novo sistema de eixos que é obtido por rotação do sistema inicial constituído por $X_1, ..., X_p$. Os novos eixos representam as direcções com máxima variabilidade (ver figura 3.1).

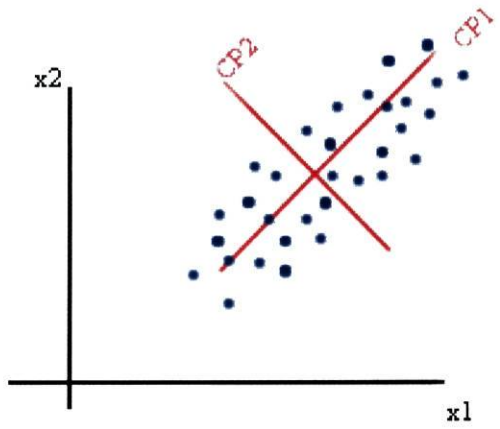


Figura 3.1: Representação geométrica das componentes principais. Os eixos originais estão a preto e as componentes principais a vermelho.

Consideremos o vector aleatório $\mathbf{X} = (X_1, \ldots, X_p)$ e a respectiva matriz de covariância Σ. A primeira componente principal é a variável

$$Y_1 = w_{11}X_1 + \ldots + w_{1p}X_p = \mathbf{w}_1^T\mathbf{X}$$

combinação linear das p variáveis originais e cuja variância é máxima. Determinar $\mathbf{w}_1$ implica a resolução de um problema de optimização[1]

$$\begin{aligned} \mathbf{w}_1 = argmax_{\mathbf{w}}\ & Var(\mathbf{w}^T\mathbf{X}) \\ \text{s.a. } & \mathbf{w}^T\mathbf{w} = 1 \end{aligned}$$

Prova-se [16, 17] que $\mathbf{w}_1$ é o vector próprio associado com o maior valor próprio da matriz de covariância Σ. Assim, a primeira componente principal é a direcção na qual

[1] $Var(\mathbf{w}_1^T\mathbf{X}) = \mathbf{w}_1^T Cov(\mathbf{X})\mathbf{w}_1 = \mathbf{w}_1^T\Sigma\mathbf{w}_1$. Note-se que $Var(\mathbf{w}_1^T\mathbf{X})$ pode ser aumentada multiplicando qualquer $\mathbf{w}$ por uma constante. A condição $\mathbf{w}_1^T\mathbf{w}_1$ permite retirar esta indeterminação e restringir o problema a vectores $\mathbf{w}_1$ unitários.

a variância da projecção é maximizada. De forma análoga, pode-se provar [16, 17] que a i-ésima componente principal é a combinação linear

$$Y_i = w_{i1}X_1 + \ldots + w_{ip}X_p = \mathbf{w}_i^T\mathbf{X}$$

que resolve o problema de optimização

$$\begin{aligned} \mathbf{w}_i = argmax_{\mathbf{w}}\ & Var(\mathbf{w}^T\mathbf{X}) \\ \text{s.a.} \quad & \mathbf{w}^T\mathbf{w} = 1 \\ & Cov(\mathbf{w}_i^T\mathbf{X}, \mathbf{w}_k^T\mathbf{X}) = 0 \quad k < i \end{aligned}$$

A direcção de projecção $\mathbf{a}_i$ é o vector próprio associado com o i-ésimo maior valor próprio da matriz Σ. Assim, as componentes principais (CP) são as combinações lineares não correlacionadas das variáveis originais $X_1, ..., X_p$ e com variância máxima. A variância explicada por cada componente é dada por $Var(\mathbf{w}_i^T\mathbf{X}) = \lambda_i$ em que λ_i é o valor próprio de Σ correspondente.
Nos casos práticos, a distribuição dos dados (e em particular μ e Σ) não é conhecida. Neste caso, dispomos de uma matriz de dados de dimensão $n \times p$ em que n é o número de observações ou amostras e p o número de variáveis. Assume-se que $\mathbf{X}$ está centrada[2]. A correspondente matriz de covariância amostral é obtida de $\mathbf{S} = \frac{1}{n-1}\mathbf{X}^T\mathbf{X}$. As CP determinam-se da mesma forma, com a diferença de que a matriz de covariância é substituída pela sua correspondente amostral, obtendo-se assim a versão amostral das CP. O número máximo de CP é igual ao número máximo de valores próprios não nulos de $\mathbf{S}$, que é igual à característica da matriz $\mathbf{X}$, $\min(p, n)$. Um dos objectivos de ACP é reduzir a dimensão dos dados. Com efeito, prova-se que a representação obtida com ACP é uma redução linear óptima da dimensão no sentido dos mínimos quadrados [16, 17]. Desta forma, escolhem-se apenas $k << p$ componentes principais, que substituem as p variáveis iniciais, obtendo-se uma redução efectiva da dimensão. Um critério para a escolha de k é a percentagem cumulativa de variância explicada pelas primeiras k componentes (habitualmente entre 85 e 95%).

O caso particular dos *microarrays* de ADN é crítico nesta formulação por duas razões amplamente discutidas neste trabalho. Por um lado, o número de variáveis é superior (largamente) ao número de amostras; por outro, esse número ronda os milhares, nos casos em que não é efectuada uma selecção de variáveis (genes) predictivas, ou as várias dezenas, nos casos em que essa selecção é efectuada. Esta particularidade acarreta problemas computacionais, como o cálculo de valores[3] e vectores próprios que se torna extremamente penoso e pouco fidedigno em termos numéricos. No entanto, estas CP podem ser obtidas de uma forma eficiente a partir da matriz $\mathbf{X}\mathbf{X}^T$ que tem dimensão $n \times n$, muito inferior à de $\mathbf{X}^T\mathbf{X}$. Seja $\mathbf{X}$ a matriz $n \times p$ ($n << p$) original da qual pretendemos determinar as CP. Assumimos que a matriz está centrada e

[2]Isto é, subtrai-se a cada variável a sua média amostral, o que implica que a soma dos elementos de cada coluna seja zero.

[3]Note-se que o número destes não nulos é n.

estandardizada de acordo com

$$x_{ij} \leftarrow \frac{x_{ij} - \bar{x}_j}{s_j\sqrt{n-1}} \tag{3.1}$$

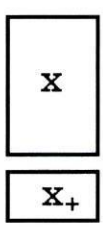
A divisão por $\sqrt{n-1}$ permite que a matriz de covariâncias exprimental possa ser obtida de $\mathbf{X}^T\mathbf{X}$, enquanto que a divisão por s_j permite reduzir o efeito das variáveis com maior dispersão, pelo que cada variável terá uma contribuição igual na determinação das proximidades entre objectos. Deste modo, $\mathbf{X}^T\mathbf{X} = \mathbf{C}$ corresponde à matriz de correlações entre variáveis. Então, se $\mathbf{w}_i, \lambda_i$ são os vectores e valores próprios respectivamente da matriz de correlação $\mathbf{C}$, estes podem ser obtidos pelas relações [17]

$$\mathbf{w}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{X}^T\mathbf{v}_i \tag{3.2}$$

em que $\mathbf{v}_i$ é vector próprio da matriz $\mathbf{X}\mathbf{X}^T$. Assim, a projecção das n observações na i-ésima componente principal é [17]

$$\mathbf{X}\mathbf{w}_i = \sqrt{\lambda_i}\mathbf{v}_i \tag{3.3}$$

Desta forma, as componentes principais são muito mais fáceis de obter, pois o cálculo de $\mathbf{v}_i$ e λ_i é muito mais simples e rápido. Assim, dependendo da situação, podemos optar pelo método mais conveniente.
E o que fazer no caso de serem disponibilizadas novas amostras? Neste caso é necessário projectá-las no novo espaço das componentes principais. O aparecimento de novas amostras corresponde a adicionar mais linhas ($\mathbf{X}_+$) à matriz de dados inicial



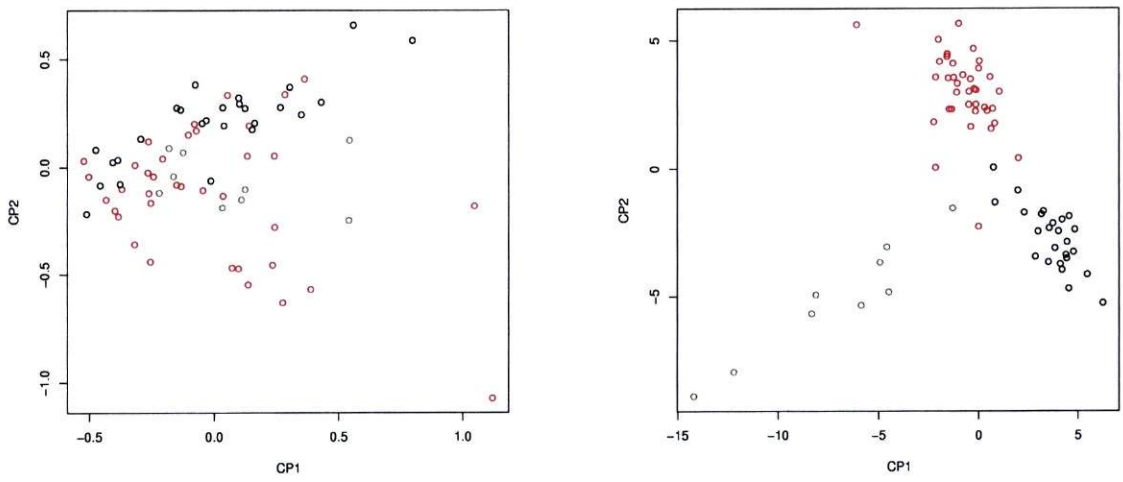
Todo o processo deverá ser efectuado em relação à nuvem de amostras inicial, pelo que a estandardização das novas amostras recorre aos parâmetros das iniciais

$$x_{+ij} \leftarrow \frac{x_{+ij} - \bar{x}_j}{s_j\sqrt{n-1}}. \tag{3.4}$$

A projecção das novas amostras na i-ésima componente principal é dada por

$$\mathbf{X}_+\mathbf{w}_i \tag{3.5}$$

onde, recorde-se, $\mathbf{w}_i$ pode ser obtido de (3.2).



(a) (b)

Figura 3.2: Projecção do conjunto LEUCEMIA nas duas primeiras componentes principais. À esquerda, foram utilizados os 3571 genes disponíveis; à direita, apenas 50 genes escolhidos com Anova foram utilizados.

A figura 3.2 mostra a projecção das 72 amostras do conjunto LEUCEMIA nas duas primeiras componentes principais. A representação refere-se às três classes ALL-B, ALL-T e AML. A figura 3.2(a) foi produzida utilizando os 3571 genes (variáveis) disponíveis; a figura 3.2(b) utiliza apenas 50 genes escolhidos com o método Anova. A diferença é evidente e mostra que a escolha de variáveis predictivas é essencial para obter representações mais discriminativas. No primeiro caso, as duas primeiras componentes principais explicam cerca de 30% da variância; no segundo caso, esse valor sobe para 55%.

3.2 Mínimos quadrados parciais

Os mínimos quadrados parciais [15, 16, 25] (*partial least squares* ou PLS) é uma técnica muito popular, em particular na área da química. O método tem uma longa e complexa história e variadas formulações. Habitualmente, tem sido expresso em termos de algoritmos para a sua implementação em regressão, o que torna difícil perceber o que realmente está a ser feito. Stone & Brooks [30] mostraram que PLS é equivalente a encontrar sucessivamente, funções lineares das variáveis predictivas $X_1, \ldots, X_p$ com covariância máxima com a variável dependente $\mathbf{y}$ (resposta), sujeito a que cada função linear seja não correlacionada com as anteriores. No caso de ACP, as componentes são construídas por forma a explicar o máximo de variação das variáveis predictivas, o que não garante que essas componentes sejam predictivas da resposta. O método PLS resolve esse problema, procurando direcções com máxima variância e máxima correlação com a resposta, propriedade importante quando falamos de predição.

O procedimento para obtenção das componentes (direcções) PLS para o caso de um problema de duas classes[4] (codificadas 0 e 1) é o seguinte. Assume-se que $\mathbf{y}$ está centrado e $\mathbf{x}_j$ estandardizado (média 0 e variância 1). O método começa por determinar o coeficiente de regressão[5] $\hat{\Phi}_{1j}$ de $\mathbf{y}$ em cada $\mathbf{x}_j$. Este é dado por $\hat{\Phi}_{1j} = < \mathbf{x}_j, \mathbf{y} >$ e a primeira direcção PLS vem $\mathbf{z}_1 = \sum \hat{\Phi}_{1j}\mathbf{x}_j$. Na verdade, $\hat{\Phi}_{1j} \propto < \mathbf{x}_j, \mathbf{y} >$, mas a constante de proporcionalidade é irrelevante para a construção da direcção. Com efeito, supondo um modelo univariado sem *intercept* da forma

$$y = \beta x + \epsilon$$

é fácil mostrar que a solução dos mínimos quadrados para β é dada por

$$\hat{\beta} = \frac{\sum x_i y_i}{\sum x_i^2} = \frac{< \mathbf{x}, \mathbf{y} >}{< \mathbf{x}, \mathbf{x} >}$$

em que $\mathbf{x} = (x_1, \ldots, x_n)^T$ é um vector de n observações da variável x e $\mathbf{y} = (y_1, \ldots, y_n)^T$ é o vector das respectivas respostas. No caso particular em estudo tem-se

$$< \mathbf{x}_j, \mathbf{x}_j >= \cos(\mathbf{x}_j, \mathbf{x}_j) \, \|\mathbf{x}_j\|^2 = \|\mathbf{x}_j\|^2 = n - 1$$

donde $\hat{\Phi}_{1j} = < \mathbf{x}_j, \mathbf{y} > /n - 1$. Assim, na construção de $\mathbf{z}_1$, as variáveis originais são pesadas de acordo com a força do seu efeito univariado em $\mathbf{y}$. O passo seguinte é ortogonalizar[6] (ou "ajustar") as variáveis originais com respeito a $\mathbf{z}_1$

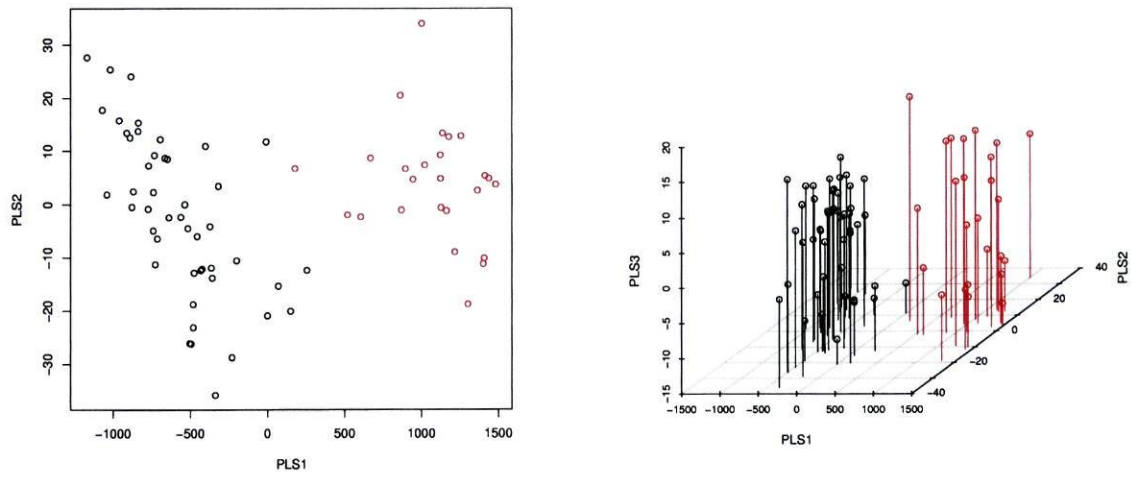
$$\mathbf{x}_j - \frac{< \mathbf{z}_1, \mathbf{x}_j >}{< \mathbf{z}_1, \mathbf{z}_1 >} \mathbf{z}_1 \qquad j = 1, \ldots, p \tag{3.6}$$

e o processo volta ao início (onde as variáveis obtidas em (3.6) desempenham o papel das variáveis originais) até serem construídas $k \leq p$ componentes PLS. O algoritmo

[4] Veja-se [24] para a extensão ao caso de múltiplas classes

[5] Modelo univariado, sem *intercept*

[6] É fácil ver que de facto as variáveis obtidas em (3.6) são ortogonais a $\mathbf{z}_1$



(a) (b)

Figura 3.3: Projecção do conjunto LEUCEMIA nas primeiras componentes PLS. À esquerda, projecção nas duas primeiras componentes; à direita, projecção nas três primeiras componentes.

completo utilizado em regressão encontra-se no Apêndice A, onde se marcam com asterisco(*) os passos necessários para se obter apenas as componentes PLS.

A figura 3.3 mostra as projecções do conjunto LEUCEMIA nas duas (figura 3.3(a)) e três (figura 3.3(b)) primeiras componentes PLS, para a distinção das classes ALL e AML. Para produzir a figura, foram seleccionados com a estatística t (2.4) apenas 50 genes predictivos. A separação das duas classes é evidente, o que revela as capacidades do método e as vantagens em incluir, na construção das componentes, informação sobre a resposta.

Analogamente a ACP, o problema da determinação das componentes PLS pode ser descrito algebricamente como um problema de optimização. A i-ésima direcção PLS resolve o problema

$$\begin{aligned} \mathbf{w}_i = argmax_{\mathbf{w}}\ & Corr^2(\mathbf{w}^T\mathbf{X}, \mathbf{y}) Var(\mathbf{w}^T\mathbf{X}) \\ \text{s.a.} \quad & \mathbf{w}^T\mathbf{w} = 1 \\ & Cov(\mathbf{w}_i^T\mathbf{X}, \mathbf{w}_k^T\mathbf{X}) = 0 \quad k < i \end{aligned}$$

O número k ($k << p$) de componentes a escolher dependerá dos dados. Como em ACP, poder-se-á escolher aquelas que explicam a maior variação de predictores e resposta ou efectuar validação cruzada e escolher k que minimiza o erro de previsão.

3.3 Variáveis canónicas

A análise discriminante procura direcções eficientes para a discriminação. Por exemplo, ACP encontra componentes úteis para a representação dos dados, mas não é garantido que essas componentes o sejam para a discriminação das diferentes classes. Desta forma, Fisher (1936, para $K = 2$ classes) e mais tarde Rao (1948, extensão para K classes) desenvolveram um método que permite obter um subespaço óptimo em termos de discriminação, com dimensão $K-1$. A ideia do método partiu do seguinte problema colocado por Fisher

Problema 1 *Encontrar a combinação linear* $y = \mathbf{w}^T\mathbf{x}$ *tal que a variância entre classes projectadas seja maximizada relativamente à variância dentro das mesmas.*

Resumidamente, procura-se a direcção de projecção tal que as classes fiquem o melhor discriminadas possível. Assim, dada uma amostra $\mathbf{x}_1, \ldots, \mathbf{x}_n$ em duas classes, a combinação linear permite obter as projecções correspondentes $y_1, \ldots, y_n$ de uma forma optimizada em termos de discriminação. A extensão para K classes é directa

Problema 2 *Encontrar as combinações lineares* $y_i = \mathbf{w}_i^T\mathbf{x}$, $i = 1, \ldots, K-1$, *tal que a variância entre classes projectadas seja maximizada relativamente à variância dentro das mesmas.*

ou seja, a cada vector $\mathbf{x}_j$ da amostra inicial, corresponde um vector $\mathbf{y}_j$, resultante da projecção daquele no subespaço óptimo $K-1$ dimensional.
É importante notar que é essencial ter em conta a variância dentro das classes e não apenas efectuar uma projecção que maximiza a distância entre os centros. Apesar de ser possível encontrar uma direcção que maximiza o mais possível esta separação, se existir uma grande sobreposição das classes (devido à natureza das covariâncias) a projecção naquele subespaço óptimo pode não ser discriminativa das classes (veja-se a figura 3.4)

Centremo-nos no caso $K = 2$ e vejamos como determinar a direcção $\mathbf{w}$. Uma forma de medir a separação ou variância entre classes projectadas é a diferença entre as projecções dos centros de cada classe, $\tilde{m}_1$ e $\tilde{m}_2$. Por seu lado, a variância ou separação dentro das classes projectadas pode ser obtida a partir das dispersões $\tilde{s}_i^2$ dentro de cada classe. Assim, o critério para o **Problema 1** pode ser escrito na forma [8]

$$\max_{\mathbf{w}} J(\mathbf{w}) = \frac{|\tilde{m}_1 - \tilde{m}_2|^2}{\tilde{s}_1^2 + \tilde{s}_2^2} \tag{3.7}$$

e, portanto, pretendemos determinar a direcção $\mathbf{w}$ que maximiza $J(.)$. É fácil ver que a expressão (3.7) pode ser escrita em termos de $\mathbf{w}$ da seguinte forma [8, 15]

$$J(\mathbf{w}) = \frac{\mathbf{w}^T S_B \mathbf{w}}{\mathbf{w}^T S_W \mathbf{w}} \tag{3.8}$$

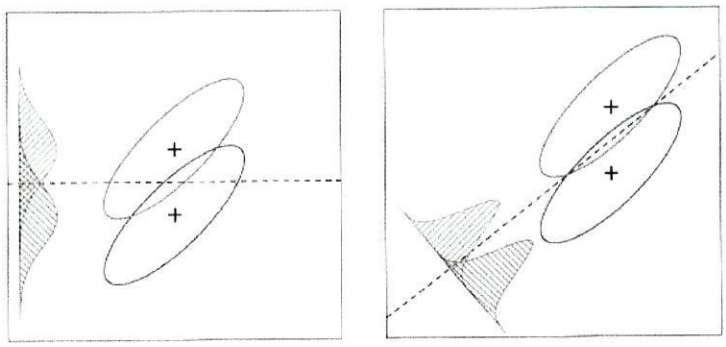


Figura 3.4: À esquerda, a linha que une os centros define a direcção que melhor os separa, mas existe uma grande sobreposição na projecção devido à natureza das covariâncias. À direita, as variáveis canónicas permitem encontrar direcções que minimizam essa sobreposição. Figura adaptada de [15].

em que S_B é a matriz de dispersão entre classes e S_W é a matriz de dispersão dentro das classes das observações iniciais.
A expressão (3.8) é designada de *quociente de Rayleigh* e mostra-se que o vector $\mathbf{w}$ que maximiza $J(.)$ satisfaz

$$S_B\mathbf{w} = \lambda S_W\mathbf{w}$$

que é um problema de valores próprios generalizado. Se S_W é não singular e, portanto, possui inversa, então obtém-se o problema de valores próprios habitual

$$S_W^{-1}S_B\mathbf{w} = \lambda\mathbf{w}$$

e prova-se que a direcção $\mathbf{w}$ procurada é o vector próprio correspondente ao maior valor próprio de $S_W^{-1}S_B$ [8, 15].

O caso de K classes é praticamente idêntico, com a diferença que, agora, pretende-se determinar $K-1$ direcções de projecção. O problema pode ser escrito na forma matricial

$$\mathbf{y} = \mathbf{W}^T\mathbf{x}$$

onde as colunas de $\mathbf{W}$ representam as direcções de projecção $\mathbf{w}_i$ (veja-se **Problema 2**). Da mesma forma podemos escrever o critério $J(.)$ em relação a $\mathbf{W}$ [8]

$$J(\mathbf{W}) = \frac{|\mathbf{W}^T S_B \mathbf{W}|}{|\mathbf{W}^T S_W \mathbf{W}|}$$

A solução para este problema é relativamente simples. As colunas de $\mathbf{W}$ são os vectores próprios generalizados correspondentes aos maiores valores próprios do problema

$$S_B\mathbf{w}_i = \lambda_i S_W\mathbf{w}_i$$

Mais uma vez, se S_W^{-1} existir, obtemos um problema habitual

$$S_W^{-1}S_B\mathbf{w}_i = \lambda_i\mathbf{w}_i$$

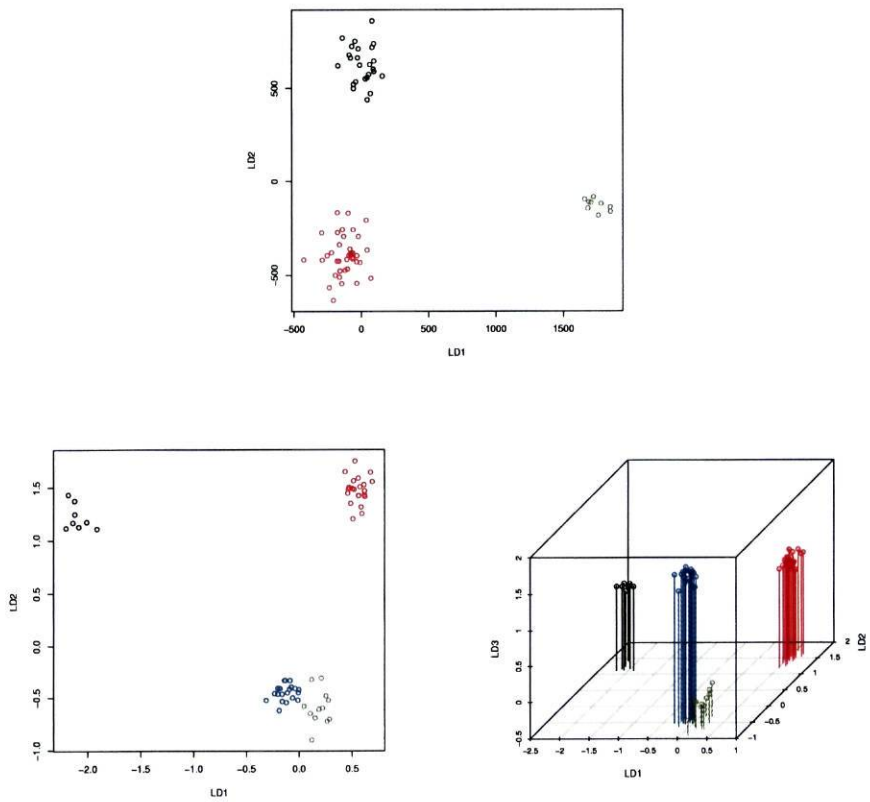


Figura 3.5: Em cima, projecção do conjunto LEUCEMIA nas duas primeiras variáveis canónicas. Em baixo, o conjunto SRBCT foi projectado nas duas primeiras (à esquerda) e nas três variáveis canónicas (à direita).

e, portanto, os $K-1$ vectores $\mathbf{w}_i$ são os vectores próprios associados com os $K-1$ maiores valores próprios de $S_W^{-1}S_B$.
Estas novas variáveis são designadas geralmente na literatura como *variáveis canónicas* e o método designa-se genericamente de LDA (*linear discriminant analysis*).
A classificação de observações utilizando variáveis canónicas pode ser efectuada de duas formas. A primeira recorre ao próprio método, que atribui a classe de cujo centro está mais próximo (no subespaço $K-1$ dimensional); a segunda, consiste em utilizar a projecção optimizada e classificar recorrendo a um outro classificador (por exemplo, k-vizinhos mais próximos).

A figura 3.5 mostra as projecções nas variáveis canónicas dos conjuntos LEUCEMIA (em cima) e SRBCT (em baixo). A discriminação entre as classes é excelente e demonstra o enorme potencial desta metodologia. No conjunto SRBCT, a projecção nas duas primeiras variáveis canónicas separa as classes, embora duas (azul e verde) poderão causar alguns problemas. Recorde-se no entanto, que o espaço óptimo tem dimensão 3 ($K-1$). O gráfico nas 3 variáveis canónicas mostra a discriminação perfeita das quatro classes.

Capítulo 4

Análise classificatória de genes

Contrariamente aos métodos de análise discriminante, os métodos de análise classificatória não requerem qualquer tipo de informação adicional para além da matriz de dados. Estes métodos são direccionados para a descoberta de padrões nos dados, sem qualquer influência de conhecimento exterior. Talvez por isso, estes métodos são na maior parte das vezes utilizados em fases exploratórias dos problemas.
O uso de métodos de análise classificatória foi uma das primeiras formas de análise de dados *microarray*. Os principais objectivos do uso destas metodologias foram fornecer formas de análise úteis para os biólogos e efectuar uma redução da enorme quantidade de informação gerada pela tecnologia *microarray* em subconjuntos mais informativos e utilizáveis; estes objectivos puderam ser alcançados utilizando os métodos de análise classificatória para, por exemplo, reordenar a lista de genes de uma forma biologicamente interpretável.
Saliente-se que estes métodos podem ser aplicados às amostras (tumores). Vários trabalhos ocupam-se deste problema, destacando-se os trabalhos de Golub *et al.* [13], que utilizaram SOMs para identificar duas classes de leucemia aguda (ALL e AML), bem como para identificar dois subtipos de ALL: células B e células T.
Um dos trabalhos de referência é também o de Alizadeh *et al.* [2], que aplicaram o método de classificação hierárquica ao conjunto de dados LINFOMA. Os autores combinaram classificação hierárquica das amostras com classificação hierárquica dos genes para identificar os genes mais importantes para a classificação das amostras. Com esta metodologia e com o conhecimento de que o tipo de tumores em estudo era clinicamente heterogéneo (40% dos pacientes têm uma resposta positiva ao tratamento, enquanto que os outros sucumbem rapidamente à doença), foi possível identificar duas formas molecularmente distintas de DLBCL[1] (*diffuse large B-cell lymphoma*).

[1] Veja-se no capítulo 5 a descrição do conjunto de dados LINFOMA.

4.1 Método de classificação hierárquica ascendente

O método mais utilizado e difundido nestes estudos é o da classificação hierárquica ascendente, para agrupar genes em grupos de semelhança. Este, permite de uma forma natural, organizar os dados *microarray*, agrupando genes com padrões de expressão semelhantes (co-expressos). Classicamente, o método começa por considerar [18] o terno

$$(O, \mu_O, d) \tag{4.1}$$

em que O é um conjunto de objectos elementares, sobre os quais se pretende determinar um sistema de classes e subclasses de associação, μ_O é uma medida positiva sobre O e que atribui um peso μ_x a cada um dos seus elementos[2] e d é um índice de distância ou semelhança sobre O. Procura-se, desta forma, representar (O, μ_O) num espaço geométrico por uma nuvem de pontos, munindo esse espaço de uma métrica tal que a distância entre pontos reflecte a semelhança entre objectos. A particularidade do método é a de passar do terno (4.1) ao terno

$$(P, \mu_P, \delta) \tag{4.2}$$

em que P é o conjunto das partes de O, μ_P uma medida positiva de pesos sobre P e δ é um índice de semelhança entre partes de O ($\delta : (P \times P, \mu_P) \rightarrow \mathbb{R}$) que será sempre função das semelhanças entre elementos de O e dos seus pesos. A partir daqui, o princípio matemático da construção da árvore de classificação hierárquica faz o resto

a cada passo, reunir os pares de classes que tornam δ mínimo

Esta construção pode ser vista como a evolução de um sistema em que o k-ésimo estado corresponde ao k-ésimo nível da árvore e que se pode representar por (T_k, μ^k). T_k é a matriz de índices δ entre classes presentes no nível k (que para o estado inicial (T_0, μ^0) é a matriz de índices entre classes singulares) e μ^k é a sucessão de pesos das mesmas. Deste modo, é intuitivo que o estado do sistema num nível k é função do nível anterior

$$(T_k, \mu^k) = f[(T_{k-1}, \mu^{k-1})]$$

com $1 \leq k \leq l-1$, em que l é o número total de níveis da árvore (o nível l corresponde à raíz).
O método de classificação hierárquica ascendente produz a cada passo uma partição π do conjunto O a classificar. Começa por considerar como partição inicial, a partição de conjuntos singulares

$$\pi_0 = \{\{x\} \,|x \in O\}$$

As agregações formam a cada passo uma nova partição, que corresponde a um novo nível da árvore. O processo termina com a partição final

$$\pi_l = \{O\}$$

[2]Esta é uma formulação mais geral. Note-se que geralmente (e o que é aqui considerado) $\mu_x = 1$ $\forall x \in O$.

correspondente à raíz da árvore. No final, é obtida uma sucessão $\pi_0, \pi_1, \ldots, \pi_l$ de partições do conjunto inicial que define os níveis da árvore.

As matrizes de dados *microarray* são numéricas, isto é, as variáveis são quantitativas. Assim o índice de semelhança entre os objectos x_i e $x_{i'}$ pode escrever-se

$$d_{ii'} = d(x_i, x_{i'}) = \sum_{j=1}^{n} d_j(x_{ij}, x_{i'j}) \tag{4.3}$$

em que n é o número de variáveis descritivas[3] e $d_j(x_{ij}, x_{i'j})$ é a contribuição da variável j para a comparação dos objectos x_i e $x_{i'}$ ($d(x_i, x_{i'})$ é a soma das contribuições das n variáveis descritivas). Desta forma, pode deduzir-se uma matriz de semelhança inicial entre os objectos a partir dos índices (4.3). Falta apenas estabelecer um critério de agregação entre classes. O critério mais utilizado na literatura designa-se de ligação média (*average linkage*) e toma como semelhança entre duas classes a distância média entre todos os pares em que cada elemento pertence a uma classe distinta

$$\delta_{AL}(G, H) = \frac{1}{N_G N_H} \sum_{i \in G} \sum_{i' \in H} d_{ii'} \tag{4.4}$$

O resultado final do algoritmo de classificação pode ser representado graficamente por uma árvore de classificação hierárquica que se denomina dendrograma (veja-se a figura 4.1 à esquerda).

Mas então que índice de semelhança escolher? As possibilidades são imensas e variadas. Uma escolha óbvia, mas não necessariamente a melhor, é a distância Euclideana

$$d_j(x_{ij}, x_{i'j}) = (x_{ij} - x_{i'j})^2$$

O índice mais utilizado na literatura (e intuitivamente mais apelativo) é o coeficiente de correlação de Pearson. Com efeito, este índice é tido como bem adaptado à noção biológica de coexpressão de dois genes, porque captura a semelhança em termos de forma e não de magnitude dos dois perfis em comparação [5, 11]. Assim, a semelhança entre dois genes i e i' é dada por

$$d(x_i, x_{i'}) = \frac{1}{n} \sum_{j=1}^{n} \left(\frac{x_{ij} - \bar{x}_i}{\sigma_{x_i}} \right) \left(\frac{x_{i'j} - \bar{x}_{i'}}{\sigma_{x_{i'}}} \right) \tag{4.5}$$

Requerendo poucas condições acerca da natureza dos dados, o método de classificação hierárquica é útil para representar variados graus de semelhança e relações mais distantes entre grupos de genes intimamente relacionados. Assim, as árvores podem ser usadas para ordenar os genes da matriz original, de tal forma que genes ou grupos de genes com padrões de expressão semelhantes são adjacentes e os comprimentos dos ramos reflectem o grau de semelhança entre objectos.

[3] Note-se que se pretende efectuar análise classificatória dos genes. As variáveis são as n amostras disponíveis.

Um dos primeiros trabalhos a utilizar esta metodologia foi conduzido por Eisen *et al.* [11]. Os autores aplicaram esta metodologia a dois conjuntos de dados obtidos por cADN. Um dos conjuntos contém os níveis de expressão da levedura *Saccharomyces cerevisiae* em vários processos celulares (como por exemplo a divisão mitótica da célula); o outro conjunto de dados contém os níveis de expressão em células humanas (para mais detalhes acerca deste último veja-se a legenda da figura 4.1; para mais detalhes de ambos veja-se [11]). Para o caso dos dados humanos obtiveram os resultados que constam da figura 4.1.

Uma propriedade marcante na figura é a presença de grandes sequências coloridas

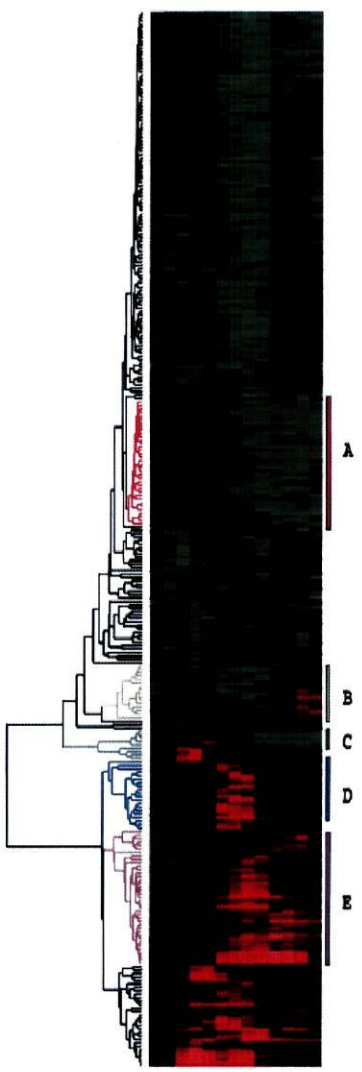


Figura 4.1: Níveis de expressão medidos ao longo do tempo de um processo de estimulação por soro de células humanas. As células foram desenvolvidas em cultura privadas de soro durante 48 horas. Posteriormente, o soro foi adicionado e foram retiradas amostras em instantes de tempo entre 0 e 24 hr, e medidos os níveis de expressão de cerca de 8600 genes distintos. À esquerda temos a correspondente àrvore de classificação hierárquica. Figura adaptada de [11].

que representam grupos de genes que partilham padrões de expressão semelhantes ao longo de diferentes condições. Estudos de randomização permitem comprovar que esta

estrutura tem origem e interpretação biológica e não é apenas uma coincidência dos dados [11]. Uma particularidade deste tipo de abordagem é a possibilidade de identificar padrões de interesse e com facilidade centrar-mo-nos mais detalhadamente nesses padrões e nos genes que para eles contribuem. O resultado talvez mais importante e admirável é obtido na análise dos maiores grupos de genes. Com efeito, observa-se uma forte tendência para que estes genes partilhem papéis em processos celulares. Aliás, uma das aplicações que se tornou bastante comum na literatura foi a utilização de classificação hierárquica para identificar grupos de genes co-regulados ou grupos funcionais, permitindo a classificação de genes de acordo com a sua função. Assim, estes métodos permitem caracterizar genes que ainda não tinham sido estudados e identificá-los com determinados processos celulares. Por exemplo, na figura 4.1 é possível identificar grupos de genes que participam na biossíntese do colesterol (grupo A) ou nos processos de renovação de tecidos e cura de feridas (grupo E).

Uma das questões bastante importantes nestes estudos é a da validação dos grupos formados. Actualmente, muita da literatura existente ocupa-se deste problema. As sugestões são variadas e vão desde a utilização de modelos de misturas de Normais [12], ao uso de métodos como ACP e LDA [20] entre outros [7, 13].

Métodos de classificação não hierárquica

Os métodos de classificação não hierárquica são também aplicados a estudos *microarray*. Os mais aplicados são os de particionamento como k-médias [26] ou os SOM (*self organizing maps*) [13, 31]. Na literatura é dado particular ênfase aos SOM pelas suas propriedades. Este método pode ser visto como uma versão restrita do método das k-médias [15]. Os SOM foram amplamente estudados e testados numa grande variedade de problemas e provaram ser significativamente superiores a métodos hierárquicos. Aliás, para o caso de dados *microarray* existem algumas críticas na utilização de métodos hierárquicos [31]: não são construídos de maneira a reflectir as várias formas nas quais os padrões de expressão podem ser semelhantes; este problema pode ser agravado com o aumento do tamanho e complexidade do conjunto de dados; possuem uma estrutura rígida, podendo agrupar observações com base em decisões locais sem a possibilidade de reavaliar o agrupamento.
Tamayo *et al.* [31] aplicaram os SOM a alguns estudos *microarray*, entre os quais ao ciclo celular da levedura *Sacharomyces cerevisiae*, tendo identificado grupos de genes co-regulados e grupos funcionais.

4.2 Agrupamento supervisionado de genes

Apesar do número elevado de genes possíveis de monitorizar com a tecnologia *microarray*, sabe-se que apenas alguns grupos de genes determinam o tipo de tecido (tumor). Como já foi referido, a identificação desses grupos é crucial tanto para o diagnóstico oncológico, como para a compreensão do funcionamento do genoma.
Existem várias formas de identificar grupos de genes semelhantes (análise classificatória), como os métodos de classificação hierárquica ou os métodos de particionamento (SOM, k-médias). No entanto, estes métodos apenas agrupam os genes pelo seu grau de semelhança, não incluindo qualquer informação sobre a variável resposta (tipo de tumor). Além disso, muitos dos métodos implicam o conhecimento de alguma forma de estrutura para os dados, o que nem sempre é possível saber na prática real. Torna-se então necessário incluir alguma informação adicional acerca dos dados (análise discriminante) por forma a construir grupos de genes predictivos da resposta.
Desta forma, Dettling *et al.* [6] propõem um método que engloba as características de ambos os métodos discriminante e de análise classificatória, por forma a construir grupos de genes com poder predictivo do tipo de tumor. O algoritmo agrupa genes efectuando uma procura para a frente e para trás (colocando e retirando genes, *forward and backward search*) por forma a optimizar uma função objectivo que avalia a capacidade discriminativa do grupo formado. Com este procedimento, obtêm grupos de 3 a 9 genes cuja expressão média permite uma boa discriminação das diferentes classes. Além disso, como os grupos contêm poucos genes, a compreensão do funcionamento do genoma fica muito mais facilitada.

4.2.1 O modelo

O modelo estocástico considerado para este problema é dado pelo par aleatório

$$(\mathbf{X}, Y) \in \mathbb{R}^p \times \Upsilon$$

em que $\mathbf{X}$ denota o perfil de expressão em p genes de uma amostra transformada logaritmicamente e normalizada (média zero e variância unitária); $\mathbb{R}$ é o conjunto dos números reais e Y é a variável resposta (tipo de tumor) que toma valores em $\Upsilon = \{1, 2, \ldots, K\}$. Consideremos em primeiro lugar o caso $K = 2$, ou seja, um problema de classificação binária.
Baseados na ideia de que apenas alguns subgrupos de genes explicam a maior (ou quase toda) parte da variação da resposta, a probabilidade condicional é modelada como

$$P(Y = 1|\mathbf{X}) = f(X_{C_1}, X_{C_2}, \ldots, X_{C_q})$$

em que $f(.)$ é uma função não linear de $\mathbb{R}^q$ em $[0, 1]$ e $\{C_1, \ldots, C_q\}$ com $q << p$ são grupos de genes disjuntos e que normalmente formam uma partição incompleta do

conjunto total[4]. X_{C_i} denota um valor de expressão representativo do grupo C_i. Este valor representativo resulta de uma simples combinação linear

$$X_{C_i} = \frac{1}{|C_i|} \sum_{g \in C_i} \alpha_g X_g \text{ com } \alpha_g \in \{-1, 1\} \tag{4.6}$$

A combinação linear (4.6) permite que um dado gene g contribua para X_{C_i} com a sua expressão trocada de sinal, $-X_g$. Isto permite tratar a sub e sobreexpressão de forma simétrica, prevenindo o caso em que genes com diferente polaridade[5] no mesmo grupo cancelem a expressão diferenciada ao efectuar a média (4.6)(veja-se a figura 4.2).

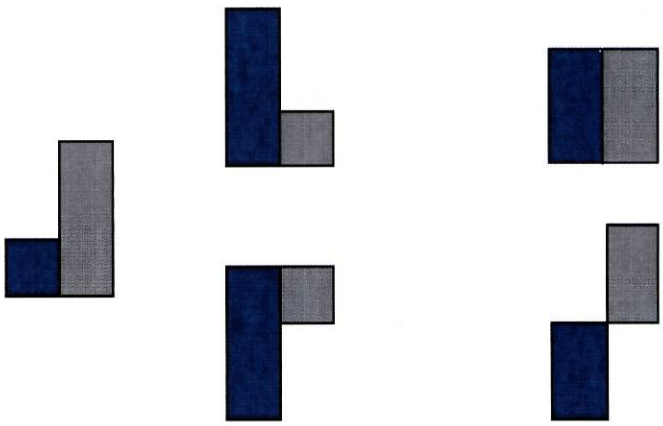


Figura 4.2: À esquerda, temos os níveis de expressão de um gene em duas classes, azul e cinza. Colocando no grupo um gene com polaridade diferente (em cima ao meio) a média provoca o cancelamento da expressão diferenciada (em cima à direita). Trocar a polaridade do segundo gene (em baixo ao meio) permite manter o potencial discriminativo do grupo (em baixo à direita).

4.2.2 Função objectivo: funções *score* e *margem*.

O processo de busca para trás e para a frente com vista a encontrar os grupos C_i, deve ser conduzido por uma função objectivo que inclua, de alguma forma, informação sobre a resposta. Consideremos agora, que dispomos de n realizações independentes e identicamente distribuídas do par aleatório $(\mathbf{X}, Y)$

$$(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)$$

com $\mathbf{x}_j \in \mathbb{R}^p$ e $y_j \in \{1, 2\}$. Recorde-se que os perfis de expressão $\mathbf{x}_j$ estão normalizados (média zero e variância unitária). A função a optimizar deverá medir de uma forma eficiente a capacidade discriminativa dos genes e/ou grupos. Para um problema de classificação binária ($K = 2$) os autores basearam-se na estatística do teste de

[4] O que não acontece com os métodos habituais de análise classificatória. Este é um ponto a favor do método proposto, pois genes que não interessam são rejeitados.

[5] Isto é, um com sobreexpressão na classe 1 e outro com sobreexpressão na classe 2.

Wilcoxon [9, 22] para duas amostras não emparelhadas. Define-se então o *score* para o gene i por

$$Score(\xi_i) = s(\xi_i) = \sum_{j \in N_1} \sum_{l \in N_2} \mathbf{1}_{[x_{ij} \geq x_{il}]} \tag{4.7}$$

em que $\xi_i = (x_{i1}, \ldots, x_{in})$ é o vector de expressões do gene i para todas as n amostras disponíveis[6], j é o índice para as N_1 amostras da classe 1 e l o índice para as N_2 amostras da classe 2.
O que a expressão (4.7) está a fazer não é mais do que, para o gene i, determinar para cada amostra da classe 1 o número de amostras da classe 2 com nível de expressão inferior e somando todos estes valores. Assim, se um dado gene tem níveis de expressão uniformemente mais baixos (altos) para a classe 1 do que para a classe 2, a função s obtém o seu valor mínimo (máximo) $s_{min} = 0$ ($s_{max} = N_1 N_2$). É precisamente neste ponto que o método inclui informação sobre a resposta (tipo de tumor).
O cálculo de s para um dado grupo de genes C_i é efectuado da mesma forma, mas neste caso usamos o perfil de expressão representativo $\xi_{C_i} = (x_{C_i,1}, \ldots, x_{C_i,n})$, em que $x_{C_i,j} = \frac{1}{|C_i|} \sum_{g \in C_i} \alpha_g \xi_{g,j}$.
Interpretar s como a estatística do teste de Wilcoxon, possibilita ordenar genes e grupos de acordo com o seu potencial discriminativo.

Recorde-se novamente o problema da existência de genes com diferentes polaridades dentro do mesmo grupo. Como ilustra a figura 4.2, existe o risco de eliminar a expressão diferenciada do grupo, o que leva a que este perca o seu potencial discriminativo. Este problema é resolvido efectuando a simples troca de sinal

$$\tilde{\xi}_i = \alpha_i \xi_i = \begin{cases} (x_{i1}, \ldots, x_{in}) & \text{se } s(\xi_i) \leq s_{max}/2 \\ (-x_{i1}, \ldots, -x_{in}) & \text{se } s(\xi_i) > s_{max}/2 \end{cases} \tag{4.8}$$

que corresponde a considerar todos os genes como subexpressos na classe 1 e a procura é conduzida nesse sentido. Esta transformação é equivalente a tomar $\alpha_g = -1$ em (4.6) para todos os genes que tendam para a subexpressão na classe 2. É fácil ver que

$$s(\tilde{\xi}_i) = \min(s(\xi_i), s_{\max} - s(\xi_i)) \tag{4.9}$$

Pelo facto de ser uma função discreta, é normal surgirem situações em que o valor de s é igual (muitas vezes zero) para diferentes genes ou grupos. De acordo com s estes genes ou grupos têm o mesmo poder discriminativo. Para obter unicidade, os autores sugerem a inclusão no processo de uma função *margem*, contínua e real, que determina a força com que um vector $\tilde{\xi}_i$ discrimina as classes[7]

$$Margem(\xi_i) = m(\xi_i) = \min_{l \in N_2}(x_{il}) - \max_{j \in N_1}(x_{ij}). \tag{4.10}$$

A função m é positiva se e só se $\min_{l \in N_2}(x_{il}) > \max_{j \in N_1}(x_{ij})$, isto é, sse ξ_i ($\tilde{\xi}_i$) está (completamente) sobreexpresso na classe 2. Neste caso, $s = 0$ o que implica que $\tilde{\xi}_i$

[6] Corresponde a uma linha da matriz de dados.
[7] Como se de uma margem de vitória se tratasse. Vence quem tem a maior margem.

discrimina perfeitamente as classes. Em caso de empates da função s, m permite distinguir o melhor gene ou grupo. O cálculo de m para um grupo é perfeitamente análogo, havendo apenas a necessidade de considerar ξ_{C_i} no lugar de ξ_i.
A função m permite ao algoritmo distinguir o melhor gene ou grupo no caso de empates da função s. Vence aquele com maior valor de m. Assim, a função objectivo é constituída por duas componentes: a função s que determina o poder discriminativo e a função m, que em caso de empate, estabelece a unicidade da escolha de s.

O primeiro passo do algoritmo é efectuar a troca de sinal de acordo com (4.8), para evitar o cancelamento de polaridades. O processo pode começar com ou sem grupos iniciais. No primeiro caso começa por determinar o padrão de expressão representativo (4.6) do grupo; no segundo, identifica o gene que optimiza s, isto é, cujo valor de s é mínimo. A construção do grupo segue de uma forma incremental, adicionando o gene que produz o menor valor de s para o grupo aumentado, ou, em caso de empates, a maior margem m. O processo é repetido até que a introdução de qualquer gene não melhora a função objectivo. Entra-se então no processo inverso, para retirar os genes que foram colocados erradamente em passos anteriores. Estes são retirados um a um, sempre que a função objectivo é optimizada. O processo de introdução e retirada de genes é repetido até que o grupo estabiliza, ou seja, a função objectivo não possa ser melhorada. Se a pretensão é formar mais do que um grupo, simplesmente retira-se do conjunto inicial o grupo de genes já formado e repete-se o processo.
Uma descrição mais detalhada do algoritmo encontra-se no Apêndice B.

4.2.3 Generalização para problemas multiclasse

O que fazer em situações em que o problema possui mais do que duas classes ou existem dentro da mesma classe subtipos que interessa identificar? A sugestão dos autores para o caso multiclasse é aplicar o procedimento *um contra todos*, reduzindo-se assim a K problemas binários (veja-se a secção 2.3.1). A cada passo (K no total) obtêm-se q grupos que discriminam cada classe relativamente às outras.

4.2.4 Robustez dos resultados. Potencial predictivo.

Para avaliar se os resultados obtidos eram de facto relevantes e não apenas uma particularidade dos dados, Dettling *et al.* [6] efectuaram um teste de permutação aleatória. Este consiste em considerar uma permutação $(y_1^*, \ldots, y_n^*)$ do vector de respostas do conjunto de dados[8] e aplicar o seu algoritmo para a construção de apenas um grupo ($q = 1$). A análise da distribuição empírica das funções s e m obtida com os dados permutados permite avaliar se os grupos construídos com os dados originais são de melhor qualidade do que o esperado. É, aliás, esta a conclusão retirada. A função m

[8]Ou seja, considerar o conjunto de dados com a classe de cada amostra possivelmente trocada.

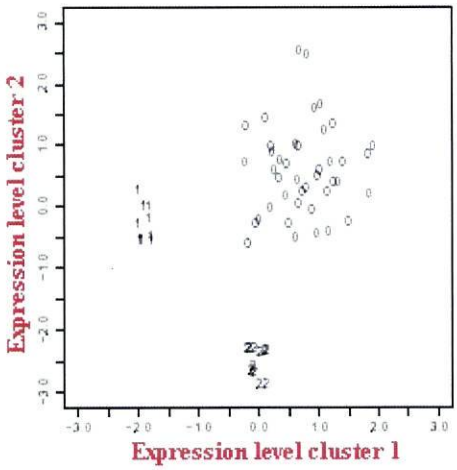


Figura 4.3: Representação bidimensional do conjunto LINFOMA. O eixo dos xx representa a expressão média do grupo formado na separação da classe 1 (FL) *versus* classes 0&2 (DLBCL&B-CLL); O eixo dos yy representa a expressão média do grupo formado na separação da classe 2 *versus* classes 0&1. Figura adaptada de [6].

possui valores muito mais positivos para os conjuntos originais (variando no intervalo $[.05, 2.52]$) do que para os dados permutados (o valor mediano para as 1000 simulações varia no intervalo [-0.12,0.12]), o que indica um menor potencial discriminativo destes. Mais, os grupos obtidos com os dados permutados possuem em média muito mais genes que os obtidos com os dados originais. A figura (4.3) demonstra as qualidades do método, onde é possível observar uma separação clara das classes.

Para avaliar o potencial predictivo dos grupos de genes obtidos com o seu algoritmo, Dettling *et al.* [6] aplicaram-no a oito conjuntos de dados reais, usando como classificador o método dos k-vizinhos mais próximos [8, 15]. Dada uma nova amostra, o método procura no conjunto de treino as k amostras que lhe são mais próximas (no sentido da distância Euclideana) e a classe maioritária neste conjunto de amostras treino é a atribuída à nova amostra. Os autores consideraram $k = 1$ como uma escolha apropriada. Os resultados para três conjuntos de dados são apresentados no capítulo seguinte.

Capítulo 5

Descrição de resultados

Após a abordagem teórica de diferentes metodologias para a análise de *microarrays* de ADN, é essencial verificar na prática o comportamento desses métodos e aferir se de facto a tecnologia *microarray* será uma ferramenta essencial para a sistematização do diagnóstico oncológico.
Este capítulo divide-se em três partes. Na primeira é efectuada uma descrição dos conjuntos de dados considerados neste trabalho. Segue-se a descrição dos resultados obtidos por algumas referências pesquisadas e uma comparação entre os vários métodos. Finalmente, a terceira parte é dedicada a experiências efectuadas por mim.

5.1 Descrição dos conjuntos de dados

Existem actualmente cerca de 8 a 10 conjuntos de dados extremamente difundidos. Neste trabalho optou-se por escolher três. Os critérios para esta selecção foram essencialmente a facilidade de obtenção dos dados e a maior frequência de utilização na literatura.

LEUCEMIA
Este conjunto de dados contém os níveis de expressão de $p = 7129$ genes em 72 pacientes sofrendo de dois tipos de leucemia: leucemia linfocítica aguda (ALL) e leucemia mielógena aguda (AML). A classe ALL pode ainda ser dividida em duas subclasses: células B (ALL-B) e células T (ALL-T). Os dados foram obtidos por hibridização de oligonucleótidos (*Affymetrix*) e encontram-se divididos em treino (38 casos) e teste (34 casos). O conjunto de treino é constituído por 27 casos ALL (19 ALL-B e 8 ALL-T) e 11 AML, todos adultos e provenientes de amostras de medula óssea. O conjunto de teste é mais heterogéneo, sendo constituído por 20 casos de crianças com diagnóstico ALL (19 ALL-B e 1 ALL-T) e 14 casos AML, 4 dos quais provenientes de adultos. Também de referir que 24 destas amostras são provenientes de medula óssea e 10 de sangue periférico(para uma descrição mais detalhada veja-se

[13]). Os dados podem ser obtidos de http://www.genome.wi.mit.edu/MPR ou de uma biblioteca especificamente criada em R e denominada **golubEsets**.

LINFOMA
Este conjunto de dados contém os níveis de expressão dos três casos mais prevalentes em adultos de tumores linfáticos: B-CLL (leucemia linfocítica crónica de células B), FL (linfoma folicular) e DLBCL (linfoma difuso de células B grandes). Os níveis de expressão foram obtidos usando um *microarray* cADN especial, designado **Lymphochip**, que contém genes expressos preferencialmente em células linfáticas ou de reconhecida importância imunológica ou oncológica. As diferentes referências fazem uso de subconjuntos destes dados, variando tanto no número de genes considerados como no número de casos em cada classe. Assim, opta-se por referir o conjunto utilizado na análise dos resultados de cada referência.
Os dados podem ser obtidos de http://genome-www.stanford.edu/lymphoma

SRBCT
Este conjunto de dados contém os perfis de expressão de $p = 2308$ genes obtidos por cADN, em amostras de tumores de células azuis pequenas e redondas (*small round blue cell tumors*) em crianças, que se dividem em 4 classes: NB (neuroblastoma), RMS (Rabdomiosarcoma), BL (linfoma de Burkitt) e EWS (sarcoma de Ewing). O conjunto encontra-se dividido em 63 amostras para treino (12 NB, 20 RMS, 8 BL e 23 EWS) e 20 para teste (6 NB, 5 RMS, 3 BL, 6 EWS e 5 não SRBCT).
Os dados podem ser obtidos de http://www.nhgri.nih.gov/DIR/Microarray.

5.2 Análise e comparação de resultados

5.2.1 Análise do conjunto Leucemia

Este conjunto de dados é dos mais utilizados nas diferentes referências. A primeira utilização destes dados foi efectuada por Golub *et al.* [13]. Após a escolha selectiva de genes (50 recorde-se), Golub *et al.* [13] prosseguiram na construção do seu predictor. Este consiste num esquema de votação pesada em que cada gene, perante uma nova amostra, vota numa classe. A classe vencedora é a atribuída. Consideremos um gene g (entre os 50 seleccionados) e x_g a expressão desse gene numa nova amostra $X = (x_1, \ldots, x_g, \ldots, x_{50})$ a classificar. Seja $\tilde{x}_g = log_{10} x_g$ e $\tilde{g} = (log_{10}(g_1), \ldots, log_{10}(g_n))$ em que g_i é a expressão de g na i-ésima amostra ($i = 1, \ldots, n$) do conjunto de treino. Defina-se

$$\begin{aligned} \tilde{g}_{norm} &= \left(\frac{\tilde{g}_1 - \mu}{\sigma}, \ldots, \frac{\tilde{g}_n - \mu}{\sigma} \right) \\ \tilde{x}_{norm} &= \frac{\tilde{x}_g - \mu}{\sigma} \end{aligned}$$

em que μ e σ são a média e desvio padrão amostrais respectivamente de $\tilde{g}$ no conjunto de treino. Consideremos também, as médias de cada classe $\hat{\mu}_1$ e $\hat{\mu}_2$ definidas por

$$\hat{\mu}_i = \frac{\sum_{j \in classe_i} \tilde{g}_{norm_i}}{|classe_i|}$$

Finalmente, define-se a fronteira de decisão entre as duas classes por $b = \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}$. O voto de cada gene é dado por

$$V = peso(g) distancia(x, b) \tag{5.1}$$

em que $peso(g) = P(g, c)$ e $distancia(x, b) = (\tilde{x}_{norm} - b)$. Assim, a expressão (5.1) fica

$$V = P(g, c)(\tilde{x}_{norm} - b) = \frac{\hat{\mu}_1 - \hat{\mu}_2}{\hat{\sigma}_2 + \hat{\sigma}_2}(\tilde{x}_{norm} - \frac{\hat{\mu}_1 + \hat{\mu}_2}{2}) \tag{5.2}$$

Note-se que o peso do gene g ($peso(g)$) não é mais do que o índice (2.1) de correlação com o vector de expressão idealizada $\mathbf{c}$ (recorde-se a figura 2.2), que reflecte o poder discriminativo de g na distinção das duas classes. Assim, os genes mais discriminativos têm um voto mais pesado. Consideremos, como exemplo, que $\hat{\mu}_1 > \hat{\mu}_2$ (o gene g está sobreexpresso na classe 1), ou seja, $P(g, c) > 0$. Se a nova amostra é tipicamente da classe 1, então $\tilde{x}_{norm} - b > 0$. Logo, por (5.2), $V > 0$. Se a amostra é tipicamente da classe 2, $\tilde{x}_{norm} - b < 0$ e $V < 0$. Assim, votos positivos são somados (para todos os genes seleccionados) resultando no voto total V_1 para a classe 1, enquanto que a soma dos módulos dos votos negativos resulta no voto total V_2 para a classe 2. A classe vencedora é a atribuída, isto é, se $V_1 > V_2$ atribuímos a classe 1, caso contrário a classe 2. É fácil verificar que no caso $\hat{\mu}_2 > \hat{\mu}_1$ o esquema funciona de igual modo.
Associado ao seu predictor, os autores definem um índice PS (*prediction strength*) que permite, para cada predição, avaliar a força da mesma, como se de uma medida de qualidade se tratasse. Este índice permite salvaguardar contra potenciais erros de predição devido a margens de vitória muito pequenas, revelando-se extremamente importante em problemas em que o custo de uma classificação incorrecta é elevado, como é o caso. O índice PS é definido por

$$PS = \frac{max(V_1, V_2) - min(V_1, V_2)}{V_1 + V_2}$$

Desta forma, não são efectuadas predições no caso de PS ser inferior a um valor pré-estabelecido. Golub *et al.* [13] escolheram $PS = 0.3$ como o valor mínimo para se efectuar a predição. Os resultados obtidos constam da tabela 5.1.

O predictor obteve 100% de classificações correctas nos casos em que $PS \geq 0.3$, tanto no conjunto de treino como no conjunto de teste. Nos casos em que $PS < 0.3$, haveria 3 erros se a classificação fosse efectuada: 1 erro (em 38 casos) no conjunto de treino e 2 (em 34 casos) no conjunto de teste . Os resultados podem ser considerados muito bons, principalmente se atendermos ao facto de que o conjunto de teste é bem mais

	Treino	Teste
Classif. Correcta ($PS \geq 0.3$)	36/36	29/29
Classif. Incorrecta ($PS < 0.3$)	1/2 (0.20)	2/5 (0.27,0.15)

Tabela 5.1: Resultados obtidos pelo predictor de votação pesada de Golub *et al.* [13]. Entre parêntesis, os valores de $PS < 0.3$ para os exemplos mal classificados.

heterogéneo do que o treino. De referir, que o valor de $PS = 0.27$ corresponde à amostra 66.

Nguyen *et al.* [25] aplicaram os métodos de ACP e PLS como um segundo passo de redução de dimensão. A classificação das amostras foi efectuada recorrendo ao discriminante logístico (LD) e ao discriminante quadrático (QDA). O primeiro passo do seu procedimento foi selecionar os genes mais predictivos de acordo com a sua estatística t (2.2). Como já foi referido, os autores levaram a cabo experiências com 50, 100, 500, 1000 e 1500 genes predictivos. Com o conjunto de treino, construíram 3 componentes PLS e 3 componentes principais, posteriormente usadas na obtenção dos resultados constantes na tabela 5.2.

	Treino				**Teste**			
	LD		QDA		LD		QDA	
p^*	PLS	CP	PLS	CP	PLS	CP	PLS	CP
50	38	38	38	38	33	33	28	30
100	38	38	38	38	32	32	29	30
500	38	38	38	38	31	31	32	28
1000	38	38	38	38	31	31	31	28
1500	38	38	38	38	31	30	30	28
50 (Golub)	38	36	38	36	33	27	31	31

Tabela 5.2: Número de classificações correctas obtidas por Nguyen *et al.* [25] para o conjunto LEUCEMIA. Para o conjunto de treino foi utilizada validação cruzada tipo *leave one out*. Em baixo, resultados obtidos usando os 50 genes predictivos obtidos por Golub *et al.* [13]

Como se pode observar, usando validação cruzada do tipo *leave one out*, obtiveram-se resultados 100% correctos para os exemplos do conjunto de treino, em qualquer das situações testadas. No caso do conjunto de teste, os resultados são extremamente satisfatórios com LD, onde há a registar, para $p^* = 50$, apenas um erro, a amostra

número 66[1]. Para um maior valor de p^* a qualidade dos resultados diminuiu, o que seria de esperar, pois são incluídos genes cada vez menos discriminativos da distinção. Com QDA os resultados não foram tão bons, com excepção do caso ($p^* = 500$,PLS) em que apenas 2 erros foram cometidos. É de salientar o facto de as componentes principais serem bastante competitivas em relação às componentes PLS, principalmente para os casos de menor valor de p^*. Aliás, para $p^* = 50, 100$ e QDA, as CP obtêm mesmo um melhor desempenho. Este facto não é de estranhar, se se pensar que a construção destas CP assenta nos genes mais discriminativos da distinção de classes e portanto a variação das variáveis predictivas já contém muita informação da variação da resposta. Foi também efectuado um estudo comparativo com o método anterior, tomando os 50 genes obtidos por Golub *et al.* [13] e aplicando a sua metodologia de 3 componentes CP e PLS e classificação efectuada com LD e QDA. Os resultados constam da tabela 5.2, última linha. As componentes PLS mostraram-se mais uma vez "imbatíveis", principalmente com o discriminante logístico. No conjunto de treino é de referir os dois erros cometidos com CP (o que não aconteceu com a sua metodologia de escolha de genes). Já sabemos que a inclusão de genes menos predictivos diminui a performance das CP. Isto leva a questionar a qualidade da métrica de correlação (2.1) proposta por Golub *et al.* [13] como critério de selecção de genes predictivos. Certamente o subconjunto de genes obtidos não é tão predictivo quanto o obtido por Nguyen *et al.* [25]. Pode-se então concluir que a escolha do índice apropriado para selecção de genes é de extrema importância. Estes factos são ainda mais reforçados pelos resultados obtidos no conjunto de teste, onde se salientam os resultados com LD em que apenas um erro é cometido com PLS (a tal amostra 66), enquanto que sete erros são cometidos com CP.

Tibshirani *et al.* [33] aplicaram o método do centróide encolhido a este conjunto de dados. Na figura 5.1 podemos observar o comportamento do método com validação cruzada (10-VC) e com o conjunto de teste. O erro mínimo com VC é atingido perto de $\Delta = 1.4$, mas utiliza cerca de 1000 genes. Esta pode ser tomada como uma solução de qualidade, mas os autores preferem escolher o valor $\Delta = 4.06$ que corresponde ao ponto a partir do qual o erro de VC começa a aumentar rapidamente. Desta forma, conseguem obter um subconjunto muito mais reduzido e interpretável de apenas 21 genes activos. Os resultados obtidos constam da tabela 5.3 e mostram que o método

Δ	Erro Treino 10-VC	Erro Teste	genes activos
1.4	0/38	2/34	1000
4.06	1/38	2/34	21

Tabela 5.3: Resultados para validação cruzada e teste, obtidos pelo método do centróide encolhido. À direita, o número de genes activos para os correspondentes valores de Δ.

[1]Esta amostra foi mal cassificada por todos os participantes da CAMDA'00 que analizaram este conjunto de dados. Não se sabe se a amostra foi correctamente etiquetada.

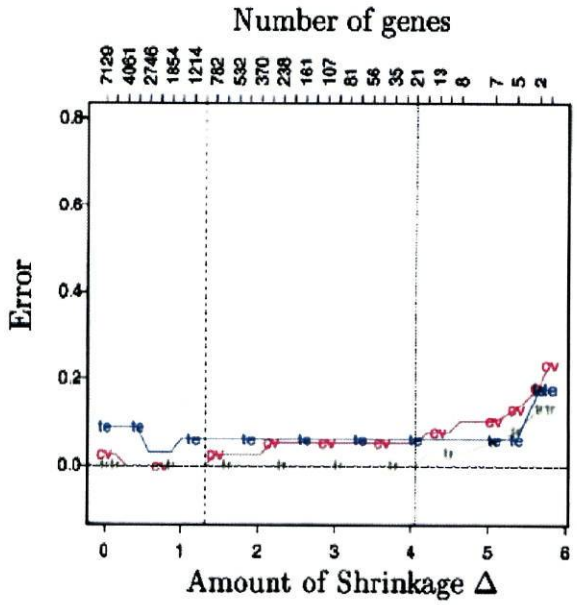


Figura 5.1: Erros de validação cruzada (vermelho) e teste (azul), para o conjuntos LEUCEMIA obtidos pelo método do centróide encolhido. Figura adaptada de [33]

comete apenas um erro no conjunto de treino (10-VC) e 2 no conjunto de teste. Repare-se que a taxa de erro obtida é igual à obtida por Golub *et al.* [13], sendo que o número de genes utilizado na classificação é muito menor (de 50 para 21), o que reforça a ideia de que aquele subconjunto de 50 genes será menos predictivo.

Dettling *et al.* [6] aplicaram o seu método ao conjunto LEUCEMIA obtendo os resultados constantes da tabela 5.4. A classificação foi efectuada recorrendo a validação cruzada do tipo *leave one out* e ao classificador do vizinho mais próximo, sobre todo o conjunto de 72 amostras, fazendo variar o número q de grupos construídos. Verifica-

	q=1	q=2	q=3	q=5	q=10	q=15	q=20
% erro	5.56	5.56	4.17	2.78	2.78	2.78	2.78
nr. erros	4/72	4/72	3/72	2/72	2/72	2/72	2/72

Tabela 5.4: Percentagem de erro e número de erros obtidos por *leave one out* sobre o total de 72 amostras, obtidos pelo método de agrupamento supervisionado para vários valores de q.

se que o aumento de q até 5 conduz à diminuição do erro, havendo uma posterior estagnação. Sabe-se que as taxas de erro obtidas por *leave one out* têm pouco viés mas uma grande variância. Neste sentido, os autores efectuaram 100 divisões aleatórias do conjunto inicial em conjunto de treino (2/3) e conjunto de teste (1/3), tendo o cuidado de colocar iguais proporções de classes em cada um. A estimativa da taxa de erro será a média das taxas obtidas com aplicação do método a cada divisão aleatória. Os resultados constam da tabela 5.5; pode observar-se que os erros obtidos

são ligeiramente superiores. Além disso, verifica-se que o aumento de q leva a uma

100 div. Aleatórias	q=1	q=2	q=3	q=5	q=10	q=15	q=20
com expressão média	6.58	4.62	4.21	3.75	3.33	3.38	3.25
com genes originais	6.33	4.79	4.50	4.08	3.67	3.75	3.79

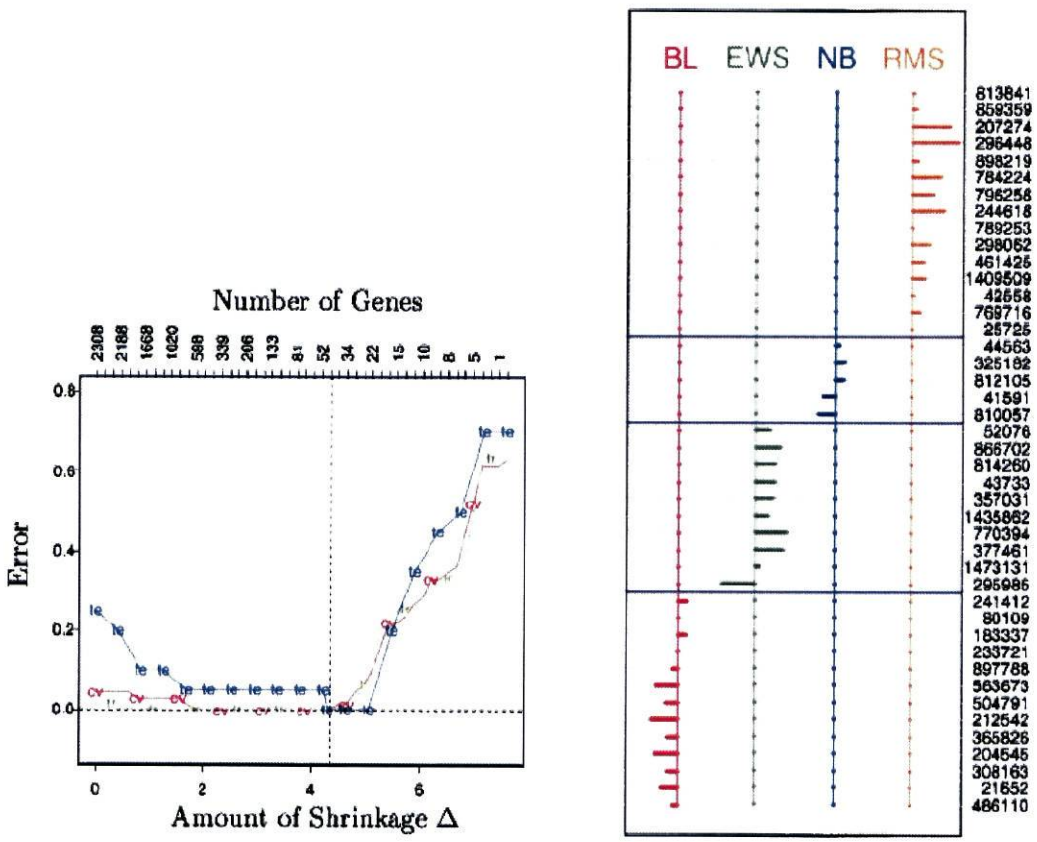
Tabela 5.5: Percentagens médias de erro obtidas a partir de 100 divisões aleatórias do conjunto inicial em treino e teste, obtidos pelo método de agrupamento supervisionado para vários valores de q.

diminuição do erro. No entanto, deve notar-se que o aumento excessivo de q pode provocar sobreajustamento. Assim, na escolha do melhor q deve ser tomado em conta um compromisso entre o erro cometido e a complexidade do modelo (e consequente sobreajustamento). É fácil ver que a partir de $q = 5$ o número médio de erros em cada divisão aleatória é inferior a 1. Recorrendo também à tabela 5.4, pode aceitar-se $q = 5$ como o número óptimo de grupos.
É interessante verificar o que acontece se em vez de considerar a expressão média dos grupos, tomar os genes obtidos de forma individual. Na tabela 5.5, podemos verificar a percentagem de erro obtida neste caso, aplicando o procedimento anterior das 100 divisões aleatórias. O erro é maior, o que leva a concluir que é vantajoso utilizar a expressão média dos grupos.

Comparação
O método de Golub *et al.* [13] tem uma vantagem óbvia com o índice PS; em caso de "dúvida"o predictor não efectua classificação, salvaguardando assim a ocorrência de erros (o que não significa que salvaguarde sempre). Por outro lado, pode tornar-se uma desvantagem porque um valor de referência PS alto pode implicar que alguns exemplos que seriam bem classificados, sejam rejeitados. De qualquer modo, esta estratégia defensiva pode certamente ser aplicada a qualquer um dos outros métodos, por exemplo, com o cálculo (estimativa) das probabilidades *a posteriori* [25, 33]. Neste sentido, parece aceitável considerar o método proposto por Golub *et al.* [13] "menos bom", nomeadamente na selecção de genes (variáveis) predictivos.
O método das componentes PLS e CP revelou-se muito bom com o discriminante logístico (LD), tanto no treino como no teste, principalmente para os valores mais baixos de p^*. O método do centróide encolhido é extremamente eficaz e consegue efectuar uma grande redução do número de genes que participam na classificação. Apesar da maior taxa de erro, este método poderá permitir uma melhor interpretação biológica relativamente ao anterior, já que o método PLS utiliza combinações lineares de um maior número de genes. O método de agrupamento supervisionado revelou-se o melhor em termos da dicotomia erro/interpretação. É de salientar, aliás, o enorme potencial interpretativo proporcionado por este método.

5.2.2 Análise do conjunto SRBCT

Tibshirani *et al.* [33] aplicaram também o seu método do centróide encolhido a este conjunto de dados. Na figura 5.2(a) podemos observar os erros de validação cruzada e conjunto de teste para estes dados.
Tanto o erro de VC como o de teste são minimizados perto de $\Delta = 4.34$ correspondendo a um total de apenas 43 genes activos dos 2308 iniciais. A figura 5.2(b) mostra as diferenças encolhidas d'_{ik} dos 43 genes activos, isto é, com pelo menos um d'_{ik} não nulo. É interessante verificar que os genes activos em cada classe são praticamente mutuamente exclusivos. Os resultados obtidos para o conjunto SRBCT pelo método



(a) (b)

Figura 5.2: a)Erros de validação cruzada e teste para o conjunto SRBCT. b)Diferenças encolhidas d'_{ik} para os 43 genes activos do conjunto SRBCT. Figuras adaptadas de [33]

de agrupamento supervisionado de genes constam da tabela 5.6. Mais uma vez verificam-se erros ligeiramente superiores para as 100 divisões aleatórias. O valor óptimo para q é 3 (aquele que minimiza o erro). Isto significa que com apenas 20 a 30 genes se consegue obter um erro médio de 0.43%. Novamente, os genes possuem individualmente um menor poder discriminativo do que em grupo, como atesta a última linha da tabela 5.6.

	q=1	q=2	q=3	q=5	q=10	q=15	q=20
leave one out	0.00	0.00	0.00	0.00	0.00	0.00	1.59
100 div. Aleatórias							
com expressão média	1.33	0.48	0.43	0.48	0.76	0.95	1.05
com genes originais	1.76	0.86	0.81	1.05	1.19	1.43	1.48

Tabela 5.6: Percentagens de erro obtidas por *leave one out* e a partir de 100 divisões aleatórias do conjunto inicial em treino e teste, obtidos pelo método de agrupamento supervisionado para vários valores de q.

Comparação
Ambos os métodos conseguem obter um erro nulo, isto é, a classificação perfeita para este conjunto de dados. A grande diferença é a do número de genes utilizados. O método do centróide encolhido utiliza 43 genes, enquanto que o agrupamento supervisionado não mais de 30. Além disso, este método permite colocar em interacção genes com diferentes polaridades e a interpretação biológica (que não deve nunca ser dissociada) é mais facilitada. Desta forma, a balança pende para o método de agrupamento supervisionado.

5.2.3 Análise do conjunto Linfoma

O subconjunto considerado por Nguyen *et al.* [25] contém a expressão de $p = 4227$ genes em 45 casos de DLBCL e 29 de B-CLL. O modelo foi construído usando 3 componentes CP e PLS e variando $p^* = 50, 100, 500, 1000$ genes predictivos seleccionados de acordo com (2.2). Os resultados obtidos constam da tabela 5.7 e foram obtidos efectuando validação cruzada tipo *leave one out* ao conjunto de 74 amostras. Com PLS obtiveram-se no máximo 2 erros, sendo que estes ocorriam sempre nas mesmas amostras; novamente, as CP têm um excelente desempenho com $p^* = 50$ e LD, aumentando drasticamente o erro com o aumento de p^*, reforçando mais uma vez a diferença (e a desvantagem) para PLS. No entanto, revelaram-se extremamente competitivas com QDA.

Por seu lado, Tibshirani *et al.* [32] aplicaram o método do centróide encolhido a um subconjunto contendo a expressão de $p = 4026$ genes em 59 casos de tumores linfáticos, compreendendo as 3 classes DLBCL, FL e B-CLL, divididas em treino (39 casos) e teste (20 casos). Com a sua metodologia, conseguiram reduzir de 4026 para 2938 genes activos com $\Delta = 0.918$ sem aumentar o erro de validação cruzada (veja-se a figura 5.3). No entanto, aplicando o método adaptativo obtiveram, para a mesma taxa de erro, apenas 48 genes activos (com pelo menos um d'_{ik} não nulo) com $\Delta = 4.41$ e $(\theta_1, \theta_2, \theta_3) = (1.88, 1.00, 1.52)$. Estes valores indicam que foram aplicados *thresholds* maiores às

	LD		QDA	
p^*	PLS	CP	PLS	CP
50	72	73	73	72
100	72	71	72	73
500	72	71	73	73
1000	72	70	73	73

Tabela 5.7: Número de classificações correctas obtidas com 3 componentes CP e PLS, aplicadas ao conjunto LINFOMA.

classes DLBCL e B-CLL, ou seja, estas são mais fáceis de identificar relativamente à classe FL. Um dos factores para esta diferença será o número de exemplos disponíveis da classe FL (apenas 9). Este foi o único conjunto de dados ao qual aplicaram o método adaptativo, que aliás, é exposto em [32], posteriormente a [33]. É de salientar que tanto o método não adaptativo como o adaptativo não cometem qualquer erro para os valores particulares de Δ escolhidos.

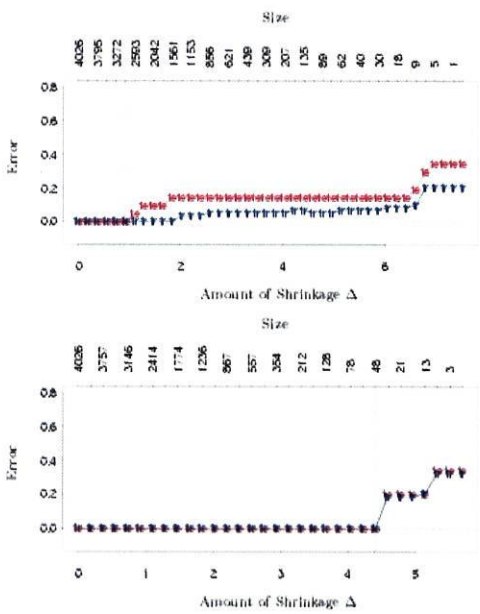


Figura 5.3: Erro de validação cruzada (tr) e erro de teste (te) com a variação do parâmetro Δ. Em cima, o método não adaptativo obtém uma solução com $\Delta = 0.918$ e 2938 genes activos. Em baixo, o método adaptativo obtém uma solução com $\Delta = 4.41$ e somente 48 genes activos, com a mesma taxa de erro que em cima. Figura adaptada de [32].

O conjunto Linfoma considerado por Dettling *et al.* [6] é constituído por 62 amostras distribuídas da seguinte forma: 42 DLBCL, 9 FL e 11 casos de CLL. Os resultados obtidos constam da tabela 5.8 e atestam mais uma vez, não só a qualidade dos grupos, bem como a vantagem em considerar a expressão média daqueles em detrimento do

poder discriminativo individual de cada gene. Note-se a classificação perfeita para $q = 10$.

	q=1	q=2	q=3	q=5	q=10	q=15	q=20
leave one out	3.23	1.61	1.61	1.61	0.00	0.00	0.00
100 div. Aleatórias							
com expressão média	2.15	2.20	1.50	0.85	0.65	0.50	0.50
com genes originais	2.43	2.29	1.76	1.05	0.81	0.81	0.86

Tabela 5.8: Percentagens de erro obtidas por *leave one out* e a partir de 100 divisões aleatórias do conjunto inicial em treino e teste, obtidos pelo método de agrupamento supervisionado para vários valores de q.

Comparação

As comparações entre os dois métodos aplicados a este conjunto de dados devem ser efectuadas com algum cuidado, pois o subconjunto considerado em cada estudo é diferente. Saliente-se que apenas os métodos do centróide encolhido e agrupamento supervisionado obtêm a classificação perfeita. Já as componentes PLS e CP cometem erros, sendo no entanto possível que o subconjunto considerado seja mais heterogéneo dentro das classes consideradas, em virtude de possuir mais amostras. Não foi possível confirmar esta hipótese.

O método do centróide encolhido tem um comportamento muito bom com o método adaptativo, pois não comete erros e retém apenas 48 genes (mesmo com uma classe de apenas 9 elementos). Este método revela-se importante para a identificação de potenciais "marcadores"de cada tipo de tumor e/ou potenciais alvos para a aplicação de fármacos. O método de agrupamento de Dettling *et al.* [6] também obtém um desempenho muito bom, salientado pela classificação perfeita com $q = 10$ grupos. Relativamente ao método anterior, terá a desvantagem de, possivelmente, usar mais genes (recorde-se que segundo os autores, os grupos têm cerca de 3 a 9 genes); no entanto, tem a vantagem óbvia em termos de interpretação biológica pelas interacções que estabelece entre genes.

5.3 As minhas experiências

A primeira experiência realizada foi na tentativa de simular o trabalho desenvolvido por Golub *et al.* [13]. Neste sentido, e usando a linguagem R, procedeu-se à construção de uma função que permitisse reter os genes mais predictivos no sentido da métrica de correlação (2.1) e a uma outra função que implementasse o esquema de votação pesada (descritas em anexo). A aplicação do método foi efectuada ao conjunto LEUCEMIA disponível na biblioteca golubEsets do R. Como já foi referido na secção 2.1, o conjunto considerado para o estudo foi aquele já pré-filtrado constituído por uma matriz de expressões contendo os níveis de 3571 genes em 72 amostras (38 para treino e 34 para teste) de dois tipos de leucemia aguda, ALL e AML. Obtidos os 50 genes mais predictivos (dos quais 25 sobreexpressos em ALL e 25 em AML), foi possível verificar uma ligeira diferença em relação àqueles descritos por Golub *et al.* [13]. Em particular, detectaram-se dois genes diferentes para o caso ALL e três para o caso AML. Por exemplo, para o caso ALL, foram obtidos os genes JO5243 e M28170, enquanto que aqueles autores obtiveram os genes M29696 e M13792. Com alguma manipulação computacional, verificou-se que os genes M29696 e M13792 possuem ordem 26 e 27 respectivamente para $\max P(g,c)$, o que leva a suspeitar que pequenas diferenças numéricas estariam na base das diferenças encontradas. Com efeito, o gene JO5243 difere cerca de 0.002 ($P(g,c)$) dos genes M29696 e M13792. Contudo, esta hipótese não se aplica aos outros casos. Desta forma, ficam algumas dúvidas acerca de qualquer processamento intermédio dos dados e que não está explícito nas referências.
Neste sentido, aplicou-se o esquema de votação pesada três vezes. Em primeiro lugar com os 50 genes predictivos obtidos na experiência ($p^* = 50$); depois com 56 genes predictivos, que permitiam ter como subconjunto aqueles obtidos por Golub *et al.* [13] ($p^* = 56$); finalmente, escolhendo os 50 genes de Golub *et al.* [13] ($p^* = 50$Golub). A tabela 5.9 apresenta apenas os resultados para as amostras onde foram cometidos

	p^*=50		p^*=56		p^*=50 Golub	
#	classe	PS	classe	PS	classe	PS
54 AML	AML*	0.10	AML*	0.15	AML*	0.23
57 AML	AML*	0.21	AML*	0.25	AML*	0.22
60 AML	ALL*	0.07	ALL*	0.01	AML*	0.06
66 AML	ALL	0.34	ALL*	0.24	ALL*	0.27
67 ALL	AML*	0.06	AML*	0.15	AML*	0.15
71 ALL	ALL*	0.28	ALL	0.34	ALL	0.30

Tabela 5.9: Resultados obtidos com o esquema de votação pesada para diferentes valores de p^* (mais detalhes no texto). À esquerda, apresenta-se o índice da amostra e a verdadeira classe.

erros ou cujo valor de PS é inferior a 0.3 (indicados com *). Os resultados obtidos com

os 50 genes de Golub *et al.* [13] foram exactamente iguais aos descritos pelos autores, validando assim as funções construídas em R. Verificaram-se algumas diferenças entre $p^* = 50, 56$ e 50 Golub, principalmente a nível de PS. Para $p^* = 50$, é cometido um erro para a amostra #66, que não é colmatado pelo valor de PS.

Foi também aplicado o método do centróide encolhido ao conjunto LEUCEMIA, na tentativa de obter um melhor resultado com o método adaptativo (não efectuado por Tibshirani *et al.*[33, 32]). O método está disponível na biblioteca **pamr** do R e um manual de utilização pode ser obtido de http://www-stat.stanford.edu/~tibs/PAM/Rdist. Em primeiro lugar aplicou-se o método não adaptativo ao conjunto contendo apenas 3571 genes (após pré-filtragem), contrariamente aos autores que utilizaram os 7129 genes constantes daquele estudo *microarray*.
Os resultados obtidos foram praticamente os mesmos. A figura 5.4 mostra o erro de validação cruzada e os erros cometidos em cada classe. O valor mínimo para o erro de validação cruzada é atingido próximo de $\Delta = 4.33$ onde 23 genes se mantêm activos. Refinando um pouco mais a lista de valores Δ a utilizar, foi possível estabelecer o valor

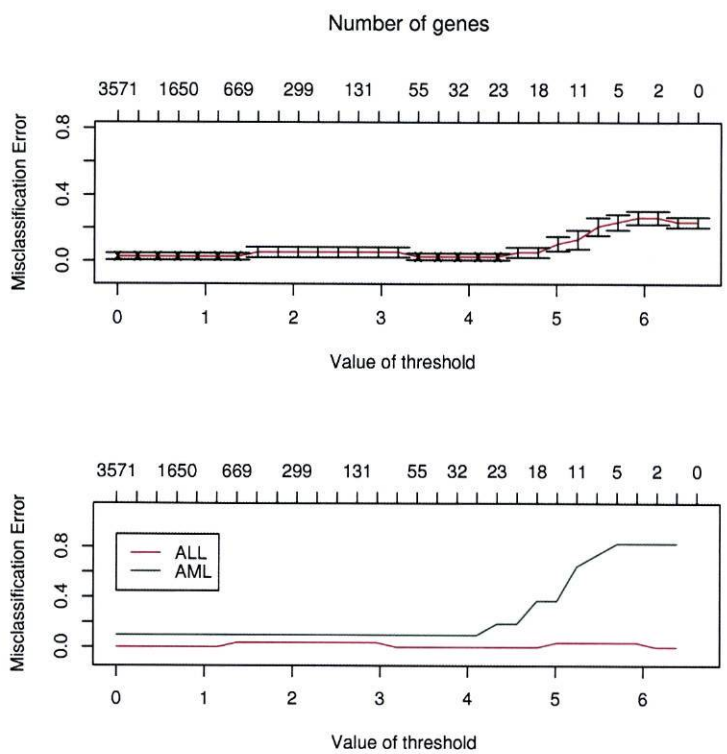


Figura 5.4: Conjunto LEUCEMIA. Em cima, erro de validação cruzada com a variação do parâmetro Δ. Em baixo, taxa de erro na validação cruzada para cada classe.

final de $\Delta = 4.313$. Com este valor, 23 genes mantêm-se activos e apenas um erro é cometido no conjunto de treino. Recorde-se que Tibshirani *et al.* [33] obtiveram um valor $\Delta = 4.06$ com apenas 21 genes activos. Estas diferenças são em parte explicadas pelo uso de validação cruzada (que depende de uma semente); também o uso de um

menor número de genes implica alterações em s_0, que obviamente afecta os valores d_{ik} (2.9). No entanto, efectuou-se também a experiência com o conjunto de dados total (7129 genes) e para o valor $\Delta = 4.06$, mais de 31 genes mantêm-se activos, o que é de estranhar. A figura 5.4 mostra que a classe AML comete mais erros. A aplicação do método adaptativo pode ter vantagens nomeadamente na redução do número de genes activos. No entanto, aplicando o método verifica-se que este não traz melhorias; em particular, o vector de escala vem $(\theta_{ALL}, \theta_{AML}) = (1, 1)$ e, portanto, o método reduz-se ao não adaptativo.
Depois do modelo ajustado partiu-se para a classificação das 34 amostras constantes do conjunto de teste. Os resultados foram semelhantes. O método comete dois erros no teste, um dos quais corresponde à amostra 66.

Os trabalhos de Golub *et al.* [13] permitiram identificar dois subtipos de ALL: células B e células T. Aplicou-se, então, o método do centróide encolhido para a classificação das três classes ALL-B, ALL-T e AML. O conjunto de treino é composto por 19 amostras ALL-B, 8 amostras ALL-T e 11 amostras AML. A figura 5.5 apresenta os resultados de validação cruzada (já com Δ refinado) e os erros de classificação em cada classe. O mínimo é atingido para $\Delta = 4.034$; com este valor, 68 genes permanecem activos e

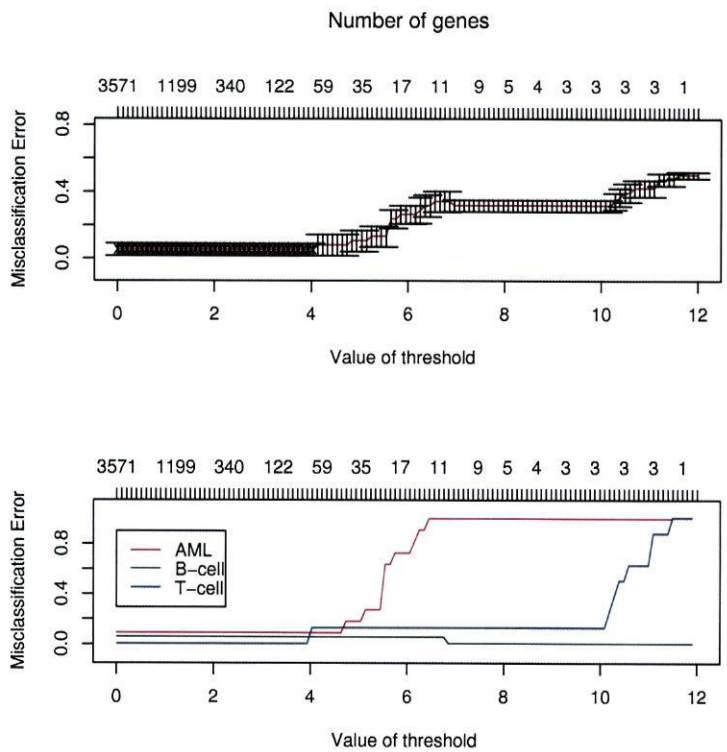


Figura 5.5: Conjunto LEUCEMIA - distinção ALL-B, ALL-T e AML. Em cima, erro de validação cruzada com a variação do parâmetro Δ. Em baixo, taxa de erro na validação cruzada para cada classe.

apenas um erro é cometido no conjunto de treino: uma amostra AML é classificada como ALL-B. Este erro foi cometido na mesma amostra que para o problema de duas

classes. Seguiu-se a aplicação do método adaptativo a este problema. Os resultados foram idênticos a anteriormente: $(\theta_{ALL-B}, \theta_{ALL-T}, \theta_{AML}) = (1, 1, 1)$, pelo que não há vantagens em utilizá-lo. Finalmente, procedeu-se à classificação do conjunto de teste. Este conjunto é composto por 19 amostras de ALL-B, 1 amostra ALL-T e 14 amostras AML. Novamente, dois erros são cometidos. A única amostra ALL-T é classificada AML, enquanto que a já esperada amostra 66 é classificada ALL-B.

O método adaptativo foi também aplicado ao conjunto SRBCT. Recorde-se que Tibshirani *et al.* [33] haviam obtido uma solução para este conjunto com $\Delta = 4.34$ e 43 genes activos, tendo utilizado apenas o método não adaptativo. Os resultados do método mostram que a classe NB é mais problemática. Com efeito, obteve-se o vector de escala $(\theta_{BL}, \theta_{EWS}, \theta_{NB}, \theta_{RMS}) = (1.372, 1.372, 1.000, 1.235)$. Com estes valores, obteve-se o erro de validação cruzada, bem como o erro de cada classe ao longo do processo (veja-se a figura 5.6). Surpreendentemente, é possível obter uma solução melhor que o método não adaptativo. O erro mínimo é atingido para $\Delta = 3.768$; o número de genes activos é 33 e não são cometidos erros. Houve, assim, uma redução de 10 genes activos.

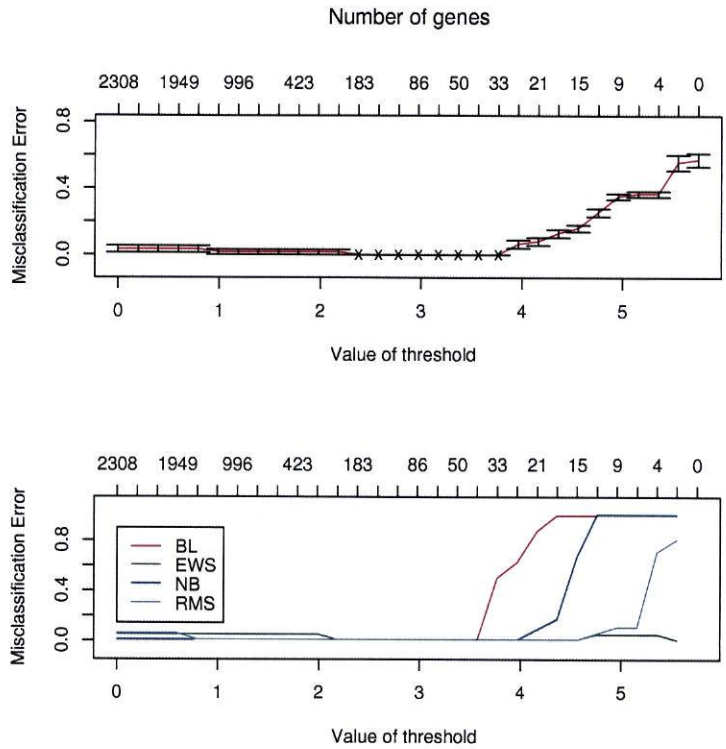


Figura 5.6: Conjunto SRBCT. Em cima, erro de validação cruzada com a variação do parâmetro Δ. Em baixo, taxa de erro na validação cruzada para cada classe.

A última experiência efectuada, prende-se com o uso das variáveis canónicas. O procedimento utilizado foi aquele que consiste em projectar as observações no espaço $K-1$ dimensional e classificar recorrendo a um classificador que não o próprio LDA. Pela sua simplicidade e bons resultados obtidos em estudos *microarray*, a escolha recaiu pelo método dos k-vizinhos mais próximos.
O primeiro passo consiste em determinar o número k de vizinhos a utilizar. Para tal, efectuou-se o seguinte: projecção das amostras nas variáveis canónicas; fazer variar k e, com *leave one out*, determinar aquele que conduz ao menor erro. As várias experiências conduzidas revelaram que $k = 1$ era uma escolha apropriada. Este resultado vem, aliás, de encontro a algumas referências [6, 10].
Neste sentido, aplicou-se o método ao conjunto LEUCEMIA para a distinção binária ALL/AML e para a distinção ALL-B/ALL-T/AML. O conjunto foi dividido em treino e teste (como já descrito anteriormente) e os resultados obtidos mostram a grande potencialidade das variáveis canónicas. Com efeito, na distinção ALL/AML, apenas um erro foi cometido no conjunto de teste, a amostra 66. Já na distinção de ALL-B/ALL-T/AML, a amostra 66 foi (surpreendentemente) correctamente classificada. Apenas um erro foi cometido: uma amostra ALL-B foi classificada em ALL-T.
Devido ao facto de apenas dispôr do conjunto de treino SRBCT, este foi dividido em treino (48 amostras) e teste (15 amostras), tentando manter as proporções de cada classe (4 recorde-se) relativamente ao conjunto inicial. Neste caso, não são cometidos quaisquer erros.
Estes resultados são um indicador da capacidade do método em discriminar classes de tumores. Assim, o método das variáveis canónicas constitui uma alternativa segura aos métodos anteriores, embora a interpretação biológica não seja tão evidente neste caso.

Capítulo 6

Conclusões e perspectivas

A principal conclusão a retirar deste estudo é que de facto a tecnologia *microarray* revela-se uma ferramenta viável para o estudo e classificação/diagnóstico oncológico. Este facto não é de estranhar, senão veja-se: o cancro resulta da multiplicação e diferenciação descontroladas de células; a multiplicação e deferenciação de uma célula depende de estruturas edificantes como as proteínas; estas estão codificadas nos genes. Assim, obter algum tipo de informação acerca do funcionamento ou dos percursos biológicos dos genes, permitirá obter informação sobre o tumor. Toda a investigação conduzida neste sentido, demonstra claramente a utilidade da tecnologia *microarray* para a sistematização do diagnóstico oncológico.

Um dos aspectos mais importantes é, obviamente, a selecção de genes (variáveis). Apesar do elevado número possível de monitorizar, pode concluir-se que apenas alguns genes, grupos de genes ou componentes (combinações lineares ou outras) são determinantes na identificação do tipo de tecido. Daí que a maior parte da atenção na literatura seja dada a este aspecto. Aliás, a selecção de genes é fundamental por três razões. Em primeiro lugar, os métodos estatísticos de classificação habituais não funcionam com um número de variáveis superior ao número de observações (o que implica a necessidade de redução do número de variáveis); em segundo lugar, de acordo com a ideia anterior, é essencial reter apenas genes predictivos da resposta (tipo de tumor); em terceiro, se a quantidade de genes retidos for reduzida, tanto melhor, pois só assim poderá ser dada uma interpretação biológica relevante, que não deve nunca ser descurada. Aliás, este último ponto torna-se ainda mais importante pelo facto de que, num futuro não muito distante, a dimensão dos dados crescerá de tal forma, por exemplo, até comportar todo o genoma humano (que se estima em cerca de 100.000 genes).

A literatura existente propõe várias alternativas para a escolha de genes. A maior parte, aplicados a situações de classificação binária, baseiam-se na ideia de um teste estatístico para determinar se um gene possui ou não poder discriminativo para a distinção das classes em estudo. Os casos multiclasse são obviamente mais difíceis e exigem, muitas das vezes, novas metodologias. Foi possível verificar que a escolha do índice para a selecção de genes deve reter especial atenção. Com o método ACP

pudémos comparar a qualidade entre os índices propostos por Golub *et al* [13] e Nguyen *et al.* [25], onde se verificou que a estatística t^* dos últimos permitia reter genes mais predictivos do que a métrica de correlação proposta pelos primeiros. O método do centróide encolhido revelou-se também muito bom, principalmente pelo número reduzido de genes que retém e pela possibilidade de tratar facilmente de problemas multiclasse; contrariamente, os anteriores necessitam de adaptações. Com o seu método de agrupamento supervisionado, Dettling *et al.* [6] obtêm excelentes resultados, chegando a superar os melhores resultados obtidos na literatura em vários conjuntos de dados. É aliás pelo trabalho destes últimos autores que se pode concluir que a interacção de genes é fundamental nestes problemas. Com efeito, a maior parte da literatura explorada separa os genes de acordo com o seu potencial discriminativo para cada classe. Por exemplo, Golub *et al.* [13] escolheram 25 genes mais predictivos da classe ALL e 25 mais predictivos da classe AML (separando assim as polaridades). Por seu lado, Dettling *et al.* [6] formam grupos passíveis de conter genes de diferentes polaridades. A interacção obtida, com a expressão média de cada grupo, revela um maior potencial predictivo do que os genes vistos como entidades discriminativas individuais. Esta particularidade, aliada ao facto de obter um conjunto de genes muito reduzido, é de extrema importância em termos de interpretação biológica.
Os resultados obtidos com os métodos ACP, PLS e variáveis canónicas também são promissores; os últimos, ao incluirem informação sobre a resposta (tipo de tumor) facilmente lidam com uma maior massa de genes. Já ACP necessita de uma pré-selecção de genes para ser competitiva. A escolha do melhor classificador para cada espaço óptimo também se revelou de extrema importância.

A tecnologia *microarray* de ADN é relativamente recente (o artigo mais antigo que disponho data de 1998) e, portanto, o seu estudo e análise são ainda prematuros. Várias formas de classificação são estudadas e aplicadas, desde as mais simples como a votação pesada, às mais elaboradas, como as redes neuronais ou máquinas de suporte vectorial [21]. Por este facto, todas as metodologias estudadas não são ainda consideradas como soluções, mas talvez como passos na direcção de um método geral de classificação para o diagnóstico oncológico. Várias são as perspectivas futuras nesta área de investigação. A construção de métodos mais robustos de selecção de variáveis e a resolução de problemas multiclasse são pontos fundamentais, bem como o estudo de conjuntos mais heterogéneos, contendo vários tipos de tumores. Um dos objectivos futuros será também usar este tipo de dados para a previsão dos tempos de sobrevivência ou da resposta a tratamentos.

Apêndice A

Algoritmo PLS

O procedimento para a obtenção dos coeficientes PLS para regressão está descrito no seguinte algoritmo. Os passos a efectuar para obter apenas as direcções de projecção estão assinalados com *.

1. * Estandardizar cada $\mathbf{x}_j$ para média zero e variância unitária. Seja $\hat{\mathbf{y}}^{(0)} = \mathbf{1}\bar{y}$ e $\mathbf{x}_j^{(0)} = \mathbf{x}_j$, $j = 1, \ldots, p$.
2. Para $m = 1, \ldots, p$
   - * $\mathbf{z}_m = \sum_{j=1}^{p} \hat{\Phi}_{mj} \mathbf{x}_j^{(m-1)}$ em que $\hat{\Phi}_{mj} = < \mathbf{x}_j^{(m-1)}, \mathbf{y} >$.
   - $\hat{\theta}_m = < \mathbf{z}_m, \mathbf{y} > / < \mathbf{z}_m, \mathbf{z}_m >$.
   - $\hat{\mathbf{y}}^{(m)} = \hat{\mathbf{y}}^{(m-1)} + \hat{\theta}_m \mathbf{z}_m$.
   - * Ortogonalizar cada $\mathbf{x}_j^{(m-1)}$ relativamente a $\mathbf{z}_m$:

$$\mathbf{x}_j^{(m)} = \mathbf{x}_j^{(m-1)} - \left[\frac{< \mathbf{z}_m, \mathbf{x}_j^{(m-1)} >}{< \mathbf{z}_m, \mathbf{z}_m >} \right] \mathbf{z}_m, \; j = 1, \ldots, p.$$

3. A sequência de coeficientes PLS nas variáveis originais $\mathbf{x}_j$ são dados por $\hat{\beta}_j^{pls}(m) = \sum_{l=1}^{m} \hat{\Phi}_{lj} \hat{\theta}_j$.

Apêndice B

Algoritmo para agrupamento supervisionado

Em seguida, apresenta-se o algoritmo (adaptado de [6]) para o agrupamento supervisionado de genes.

1. Começar com a matriz de expressões total $X_{p\times n}$ (conjunto de treino), cujas linhas representam genes e as colunas amostras de 2 tipos de tecido diferentes, normalizadas.
2. Determinar s como em (4.7) para cada gene $\xi_i = (x_{i1}, \ldots, x_{in})$, ou seja, para cada linha de X. Efectuar a troca de sinal de acordo com (4.8). Esta operação transforma o *score* para $s(\tilde{\xi}_i) = \min(s(\xi_i), s_{\max} - s(\xi_i))$.
3. a) Se não for dado nenhum grupo inicial, identificar o gene i^* com o menor $s(\tilde{\xi}_i)$. Se houver mais do que um, escolher aquele com a maior margem $m(\tilde{\xi}_i)$ como em (4.10). Estabelecer a média inicial do grupo, ξ_C, igual ao vector $\tilde{\xi}_{i^*}$ do gene escolhido.
   b) Se for fornecido um cluster C inicial, determinar a expressão média dos seus genes
   $$\xi_C = \frac{1}{|C|}\sum_{g\in C}\tilde{\xi}_g = \frac{1}{|C|}\sum_{g\in C}\alpha_g.(x_{g1}, \ldots, x_{gn})$$
4. Procura para a frente
   Calcular a média do conjunto formado pelo grupo actual e cada gene i possível de entrar
   $$\xi_{C+i} = \frac{1}{|C|+1}\left(\sum_{g\in C}\tilde{\xi}_g + \tilde{\xi}_i\right)$$
   Identificar o gene vencedor $i^* = argmin_i s(\xi_{C+i})$, isto é, aquele que conduz ao menor *score*. Se não for único, identificar o gene i^* que simultaneamente optimiza $m(\xi_{C+i})$.
5. Repetir o passo 4 até que nenhum gene permite optimizar a função objectivo.
6. Procura para trás
   Excluir cada gene i do grupo actual C separadamente e determinar a média dos restantes
   $$\xi_{C-i} = \frac{1}{|C|-1}\sum_{g\in C-\{i\}}\tilde{\xi}_g,\ i \in C$$
   Calcular s e m para cada ξ_{C-i}. Identificar (como em 4) o gene i^* cuja exclusão optimiza s, ou em caso de não ser único, optimiza simultaneamente s e m.
7. Repetir 6 até que a exclusão não produza um melhoramento da função objectivo.
8. Repetir os passos 4 a 7 até que o grupo estabilize e a função objectivo é óptima.
9. Se pretender mais do que um grupo, retirar de X os genes de C e voltar ao passo 3.

Referências

[1] The Chipping Forecast. *Supplement to Nature Genetics*, 21, 1999.

[2] A. Alizadeh, R. Eisen, M. Davis, C. Ma, I. Lossos, A. Rosenwald, J. Boldrick, H. Sabet, T. Tran, X. Yu, J. Powell, L. Yang, G. Marti, T. Moore, J. Hudson, L. Lu, D. Lewis, R. Tibshirani, G. Sherlock, W. Chan, T. Greiner, D. Weisenburger, J. Armitage, R. Warnke, R. Levy, W. Wilson, M. Grever, J. Byrd, L. Staudt, P. Brown, and D. Botstein. Distinct types of diffuse large B-cell lymphoma identified by gene expression profiling. *Nature*, 403:503–511, 2000.

[3] R.E. Bellman. *Adaptive Control Processes.* Princeton University Press, 1961.

[4] A. Brazma, H. Parkinson, T. Schlitt, and M. Shojatalab. A quick introduction to elements of biology - cells, molecules, genes, functional genomics, microarrays. *EMBL*, http://industry.ebi.ac.uk/~brazma/Biointro/biology.html:draft, Oct. 2001.

[5] A. Brazma and J. Vilo. Gene expression data analysis. *FEBS Letters*, 480:17–24, 2000.

[6] M. Dettling and P. Bühlmann. Supervised Clustering of Genes. *Genome Biology*, 3(12), 2002.

[7] E. Dougherty, J. Barrera, M. Brun, S. Kim, R. Cesar, Y. Chen, M. Bittner, and J. Trent. Inference from clustering with application to gene expression microarrays. *Journal of Computational Biology*, 9(1):105–126, 2002.

[8] R. Duda, P. Hart, and D. Stork. *Pattern Classification.* Wiley, 2nd edition, 2000.

[9] E. Dudewicz and S. Mishra. *Modern Mathematical Statistics.* Wiley, 1st edition, 1988.

[10] S. Dudoit, J. Fridlyand, and T. Speed. Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association*, 97:77–87, 2002.

[11] M. Eisen, P. Spellman, P. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. *Proc. Natl. Acad. Sci. USA*, 1998.

[12] D. Ghosh and A. Chinnaiyan. Mixture modelling of gene expression data from microarray experiments. *Bioinformatics*, 18(2):275–286, 2002.

[13] T. Golub, D. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Cagliuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: class discovery and class prediction by gene expression monitoring. *Science*, 286(5439):531–537, Oct. 1999.

[14] C. Harrington, C. Rosenow, and J. Retief. Monitoring gene expression using DNA microarrays. *Current opinion in Microbiology*, 3:285–291, 2000.

[15] T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning: data mining, inference and prediction.* Springer, 1st edition, 2001.

[16] I.T. Jolliffe. *Principal Component Analysis.* Springer, 2st edition, 2002.

[17] L. Lebart, A. Morineau, and J. Fénelon. *Traitement des donnés statistiques.* Dunod, 2nd edition, 1992.

[18] I.C. Lerman, Ph. Peter, and Leredde H. Principes et calculs de la methode implante dans le programme CHAVL I et II. *La Revue de Modulad*, I: 1993, numro 12,pp. 33-70,II: 1994, numro 13, INRIA.

[19] J.H. Maindonald. *Using R for Data Analysis and Graphics: an introduction.* http://cran.r-project.org, 2001.

[20] M. Méndez, C. Hödar, C. Vulpe, M. González, and V. Cambiazo. Discriminant analysis to evaluate clustering of gene expression data. *FEBS Letters*, 522:24–28, 2002.

[21] S. Mukherjee, P. Tamayo, D. Slonim, A. Verri, T. Golub, J. Mesirov, and T. Poggio. Support vector machine classification of microarray data. *MIT technical report*, 1998.

[22] B.J.F. Murteira. *Probabilidades e Estatística*, volume II. McGraw-Hill, 1990.

[23] R. Nadon and J. Shoemaker. Statistical issues with microarrays: processing and analysis. *TRENDS in Genetics*, 18:265–271, May 2002.

[24] D. Nguyen and M. Rocke. Multi-class cancer classification via partial least squares with gene expression profiles. *Bioinformatics*, 18:1216–1226, 2002.

[25] D. Nguyen and M. Rocke. Tumor classification by partial least squares using microarray gene expression data. *Bioinformatics*, 18:39–50, 2002.

[26] S. Raychaudhuri, P. Sutphin, J. Chang, and R. Altman. Basic microarray analysis: grouping and feature reduction. *TRENDS in Biotechnology*, 19:189–193, May 2001.

[27] B.D. Ripley and W.N. Venables. *Modern Applied Statistics with S.* Springer, 4th edition, 2002.

[28] G. Sherlock, T. Boussard, A. Kasarskis, G. Binkley, J. Matese, S. Dwight, M. Kaloper, S. Weng, H. Jin, C. Ball, M. Eisen, P. Spellman, P. Brown, D. Botstein, and J. Cherry. The Stanford Microarray Database. *Nucleic Acids Research*, 29:152–155, 2001.

[29] D. Slonim, T. Golub, P. Tamayo, J.P. Mesirov, and E.S. Lander. Class prediction and discovery using gene expression data. *RECOMB*, pages 263–272, 2000.

[30] M. Stone and R.J. Brooks. Continuum regression: cross validated sequentially constructed prediction embracing ordinary least squares, partial least squares and principal component regression (with discussion). *J. R. Statist. Soc. B*, 52:237–269, 1990.

[31] P. Tamayo, D. Slonim, J. Mesirov, Q. Zhu, S. Kitareewan, E. Dmitrovsky, T Golub, and E. Lander. Interpreting patterns of gene expression with self-organizing maps: Methods and application to hematopoietic differentiation. *Proc. Natl. Acad. Sci. USA*, 96:2907–2912, 1999.

[32] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Class prediction by nearest shrunken centroids, with applications to DNA microarrays. *Stanford technical report*, June 2002.

[33] R. Tibshirani, T. Hastie, B. Narasimhan, and G. Chu. Diagnosis of multiple cancer types by shrunken centroids of gene expression. *PNAS*, 99:6567–6572, May 2002.

[34] W.N. Venables, D.M. Smith, and R Development Core Team. *An introduction to R.*

[35] Y. Yang, S. Dudoit, P. Luu, and T. Speed. Normalization for cDNA Microarray Data. *Microarrays: Optical Technologies and Informatics, Proceedings of SPIE*, 4266:141–152, 2001.

Anexos

As três funções que se seguem permitem determinar as componentes principais a partir da matriz $\mathbf{X}\mathbf{X}^T$, conforme descrito no terceiro capítulo. A função eigen.val.vec determina os valores e vectores próprios de $\mathbf{X}\mathbf{X}^T$; por seu lado, calc.vec.pp determina os vectores próprios de $\mathbf{X}^T\mathbf{X}$ de acordo com (3.2), bem como a projecção do conjunto de treino nas componentes principais; comp.prin serve de função principal, na qual é efectuada a normalização da matriz de treino e a projecção do conjunto de teste nas componentes principais. Os parâmetros de saída são: os vectores e valores próprios, a percentagem de variância explicada por cada componente, a percentagem acumulada e as projecções do conjunto de treino e de teste nas componentes principais.

```
comp.prin<-function (treino,teste) {
treino.norm<-scale(treino)/sqrt(nrow(treino)-1)

val.vec.pp<-eigen.val.vec(treino.norm)
val.pp<-val.vec.pp$val
perc.var<-val.vec.pp$perc
perc.acumul<-val.vec.pp$acumul

U.scores<-calc.vec.pp(treino.norm,val.vec.pp$val,val.vec.pp$vec)
vec.pp<-U.scores$U
scores.treino<-U.scores$scores

teste.norm<-scale(teste,center=attr(treino.norm,"scaled:center"),
                        +scale=attr(treino.norm,"scaled:scale"))

scores.teste<-teste.norm%*%U.scores$U
return(vec.pp,val.pp,perc.var,perc.acumul,scores.treino,scores.teste)
}

eigen.val.vec<-function(matriz){
d<-dim(matriz)
m<-min(d[1],d[2])
print(m)
a<-eigen(matriz%*%t(matriz))

perc<-a$values[1:m]/sum(a$values[1:m])
acumul<-cumsum(perc)
val<-a$values[1:m]
vec<-a$vectors[,1:m]

return(val,vec,perc,acumul)
}
```

```
calc.vec.pp<-function (X,val,vec,n){

d<-dim(vec)
U<-t(X)%*%vec

for (i in 1:d[2]){

   U[,i]<-U[,i]/sqrt(val[[i]])
   }

scores<-X%*%U
return(U,scores)
}
```

As funções P.filt e classificador que se seguem permitem aplicar a metodologia proposta por Golub *et al.* [13]. A primeira tem como parâmetros de entrada o conjunto de treino, conjunto de teste e o número n de genes a reter com $P(g,c)$. A função começa por calcular $P(g,c)$ para cada gene. Após ordenaçã destes valores são escolhidos $n/2$ com $\max P(g,c)$ e $n/2$ com $\max -P(g,c)$. O parâmetro de saída corr contém os valores de $P(g,c)$ para os n genes escolhidos; treino e teste contêm os conjuntos de treino e teste com apenas os níveis de expressão dos n genes escolhidos.

```
P.filt<-function(treino,teste,n){
p<-c()
d<-dim(treino)
for (i in 1:d[1]){
a<-mean(treino[i,1:27])
b<-mean(treino[i,28:38])
s1<-sd(treino[i,1:27])
s2<-sd(treino[i,28:38])
p<-c(p,(a-b)/(s1+s2))
}
m<-complete.cases(p)
treino<-treino[m,]
teste<-teste[m,]
p<-p[m]
p1<-sort(p)
P.ALL<-p>=p1[length(p1)-n/2+1]
P.AML<-p<=p1[n/2]
treino<-rbind(treino[P.ALL,],treino[P.AML,])
teste<-rbind(teste[P.ALL,],teste[P.AML,])
corr<-c(p[P.ALL],p[P.AML])
return(corr,treino,teste)
}
```

A função classificador aplica o método de votação pesada. Os parâmetros de entrada são aqueles de saída da função P.filt. A função começa por efectuar as transformações como descritas por Golub *et al.* [13] (logaritmo,normalização). Segue-se o cálculo do voto V de cada gene, as somas para cada classe e valor PS. Os parâmetros de saída são um vector classes com as classificações de cada amostra do conjunto de teste e os correspondentes valores de PS em PS.

```
classificador<-function(treino,teste,corr){

treino<-log10(treino)
teste<-log10(teste)
d<-dim(teste)

media<-rowMeans(treino)
desvio<-sd(t(treino))
treino.norm<-t(scale(t(treino),center=media,scale=desvio))

classes<-c()
PS<-c()
for (i in 1:d[2]){

        V<-c()
        for (j in 1:d[1]){

        x.norm<-(teste[j,i]-media[j])/desvio[j]

        media1<-mean(treino.norm[j,1:27])
        media2<-mean(treino.norm[j,28:38])

        b<-(media1+media2)/2

        V<-c(V,corr[j]*(x.norm-b))
        }

        V1<-sum(V[V>0])
        V2<-sum(V[V<0])
        ps<-(max(abs(V2),V1)-min(abs(V2),V1))/(abs(V2)+V1)

        if (V1<abs(V2))
                if (ps>=0.3)
                classes<-c(classes,"AML")
                else {
                classes<-c(classes,"AML*")
                     }
        else {
                if (ps>=0.3)
                classes<-c(classes,"ALL")
                else {
                classes<-c(classes,"ALL*")
                     }
        }
        PS<-c(PS,ps)
}
return(classes,PS)
}
```

As funções seguintes permitem efectuar uma selecção de genes. A primeira, anova.filt, usa o método Anova; a segunda, t.filt, usa a estatística t. Ambas fazem uso de funções já existentes no R e de bibliotecas específicas para a filtragem de matrizes *microarray*.

```
anova.filt<-function (treino,teste,cl,p) {
Afilter<-Anova(cl,p)
aff<-filterfun(Afilter)
filt<-genefilter(treino,aff)
print(sum(filt))
teste<-teste[filt,]
treino<-treino[filt,]
return(treino,teste)
}

t.filt<-function (treino,teste,cl,p){
tf<-ttest(cl,p)
ff<-filterfun(tf)
filt<-genefilter(treino,ff)
print(sum(filt))
teste<-teste[filt,]
treino<-treino[filt,]
return(treino,teste)
}
```

A função Var.Canonicas permite obter as variáveis canónicas. Os parâmetros de entrada são o conjunto de treino, o vector de classificações deste e o conjunto de teste. Os parâmetros de saída são a projecção do conjunto de treino e do conjunto de teste nas variáveis canónicas e os vectores de projecção. É feito uso da função lda da biblioteca MASS. Agradeço à minha colega Janete Borges por me ter facultado a função.

```
Var.Canonicas<-function (treino,classe.treino,teste){
var.can<-lda(treino,classe.treino)
vc.treino<-as.matrix(treino)%*%var.can[[4]]
vc.teste<-as.matrix(teste)%*%var.can[[4]]
return(vc.treino,vc.teste,var.can)
}
```

A função knn.cvk permite, utilizando *leave one out*, determinar o melhor k para o método dos vizinhos mais próximos. O método *leave one out* é implementado pela função knn.cv da biblioteca class.

```
knn.cvk<-function (treino,cl,k=1:15,l=0,prob=FALSE,use.all=TRUE){
  cv.err<-rep(0,length(k))
  cl.pred<-matrix(NA,nrow(treino),length(k))
  for(j in (1:length(k)))
  {
   cl.pred[,j]<-knn.cv(treino,cl,k[j],l,prob,use.all)
   cv.err[j]<-sum(cl!=cl.pred[,j])
  }
  k0<-k[which.min(cv.err)]
  return(k=k0,pred=cl.pred[,which.min(cv.err)])
}
```

Finalmente, a função pls, permite obter as componentes PLS. Os parâmetros de entrada são o conjunto de treino, o vector de respostas correspondente (binário) e o número n de componentes desejadas. Os parâmetros de saída são as direcções de projecção e o conjunto de treino projectado nas n componentes PLS. Esta função baseou-se nos algoritmos propostos em [25, 15].

```
pls<-function(treino.pls,y,n){

        X<-scale(treino.pls)
        y1<-c(rep(mean(y),length(y)))
        d<-dim(X)
        Vec.pls<-matrix(0,d[2],n)
        pls.treino<-matrix(0,d[1],n)

        for (i in 1:n){

                phi<-crossprod(X,y)
                Vec.pls[,i]<-phi
                z<-X%*%phi
                pls.treino[,i]<-z

                for (j in 1:d[2]){

                a<-t(z)%*%X[,j]
                b<-vecnorm(z)^2
                X[,j]<-X[,j]-a[1,1]/b*z
                }
        }
return(Vec.pls,pls.treino)
}
```