Instituto de Matemática e Estatística
Universidade de São Paulo

Estudo comparativo de medidas de dependência e aplicações em dados de expressão gênica

Suzana de Siqueira Santos

Orientador: Prof. Dr. André Fujita.

Apoio financeiro fornecido pelo PIBIC/CNPq, processo 134653/2012-2

São Paulo, 2012

"Vivemos pelo que acreditamos.
Nosso limite está nisso.
Portanto, se cremos no que ilimitado é,
sem limites viveremos."

Allan Kardec

Agradecimentos

Ao PIBIC/CNPq pelo apoio financeiro concedido a este trabalho.

Aos autores Ruth Heller, Yair Heller e Malka Gorfine por gentilmente nos fornecerem a implementação de uma das medidas selecionadas no nosso estudo.

Ao professor Dr. André Fujita pela paciência, compreensão e grande dedicação com a qual orientou minha iniciação científica e por me proporcionar tanto incentivo e experiências gratificantes, como a monitoria aos alunos de pré-iniciação científica.

À minha família, aos meus amigos e professores sem os quais esses quatro anos de BCC não seriam tão proveitosos e tão importantes na minha vida.

A todos os amigos IMEanos por fazerem parte desta jordada BCC!

Aos professores e funcionários por fazerem muito além de seu papel e contribuirem tanto para a minha formação.

À Francisca, por sempre me ajudar com muita disposição todas as vezes que a importunei na seção de alunos.

Ao professor Dr. Fabio Kon, pela dedicada supervisão durante a monitoria de Introdução à Computação e pelas valiosas recomendações.

Ao Giuliano pela oportunidade de ser RC e pelas grandes contribuições ao BCC no IME.

Aos meus veteranos Susanna e Gustavo, por todos os inúmeros conselhos.

À minha "bixete" Ludmila, por contagiar-me de alegria, ao Pedro e a tantos outros "bixos" com os quais tive contato quando fui monitora e que me proporcionaram uma experiência tão gratificante.

Muito obrigada meus colegas ou companheiros nos intervalos IMEanos, Samu, Jackson, Ná, Felipe, Renato, Jefferson, Wil (que também é meu fornecedor de caronas), Goroba, Omar, Haruki, Coelho, Brócolis, Henrique, Wall, Jé, Tiozão, Katague, Paulo, Miojo, Manzo, Gian e Douglas pelas inúmeras gargalhadas e por tornarem a trajetória no IME tão especial.

Agradeço mais uma vez ao Jackson pela amizade sincera, as longas conversas filosóficas e por estar também ao meu lado nos momentos mais difíceis.

Reforço minhas palavras de agradecimento à pessoa querida que é o Samuel, meu namorado, grande amigo e parceiro de numerosas e divertidas conversas malucas, que multiplicou meus sonhos e alegrias, dividiu minhas tristezas, sendo um verdadeiro companheiro nessa jornada BCC e em todos os momentos. Não posso deixar de agradecer também à sua família, ao Depa, à Luiza, à Jacque, pelo carinho e apoio, e por constituirem, para mim, uma família por afinidade.

Obrigada Amanda, Nébs, Keyla e Ju, minhas amigas de colégio e para toda a vida, por tantos momentos felizes!

Obrigada, Loly, minha amiga de infância, por compartilhar comigo tantos sonhos.

Agradeço às minhas três priminhas Luiza, Julia e Mariana, fontes de alegria e muito orgulho!

À minha avó Toninha, às minhas tias Lolinha, Inês e Sandra e ao tio Chico, pela torcida, carinho, ensinamentos e por tantas vezes me compreenderem quando estive ausente.

Todo agradecimento é pouco à família que constitui meu lar e que é minha fonte de inspiração e coragem: aos meus pais Toninho (*in memoriam*) e Juliana pelo empenho na minha formação, apoio incondicional, pelos momentos de muita alegria e por sempre me proporcionarem tudo o que precisei; às minhas irmãs Ana Gabriela e Cristiana, minhas grandes companheiras de todas as horas; à minha cachorrinha Lady por sua lealdade, amizade e por me aturar durante esses quase treze anos.

Agradeço, *in memoriam*, ao meu pai Toninho, aos meus avós Alberto, Maria Georgina e José e aos meus tios Antonio e Dé, pelos inestimáveis exemplos de conduta e sabedoria e pelas lembranças incríveis que deixaram.

Dedico este último parágrafo em agradecimento a Deus, pela vida, por cada aprendizado e cada oportunidade de concretizar sonhos.

Resumo

Diversas áreas do conhecimento utilizam-se de medidas de dependência estatística para detectar associações entre variáveis de um determinado conjunto de dados. Dada a diversidade de relações possíveis entre as variáveis sob análise (linear, não linear e não funcionais), é desejável a utilização de medidas que sejam capazes de reconhecer numerosos tipos de associação. Ilustrações dessa diversidade existem na Biologia Molecular, onde há, segundo alguns estudos, relações não monotônicas, que não são identificadas pelas mais tradicionais medidas de dependência estatística, como a correlação de Pearson e Spearman. Neste trabalho, apresentamos um estudo comparativo entre as medidas de Pearson, Spearman e Kendall, a correlação de distância (Dcor) a informação mútua (IM), o coeficiente de informação máxima (CIM) e as medidas D de Hoeffding e de Heller, Heller e Gorfine (HHG), a fim de ilustrar as potencialidades e limitações de cada uma, considerando diversos tipos de relação (linear, monotônica não linear, não monotônica e não funcional), diferentes tamanhos de amostra e a presença/ausência de *outliers*. Além disso, aplicamos as medidas estudadas em dados biológicos reais advindos da tecnologia de microarranjos de DNA.

Palavras-chave: dependência; correlação; informação mútua; medida D de Hoeffding; Coeficiente de Informação Máxima; HHG.

Sumário

Parte I

Parte Objetiva

Capítulo 1
Introdução

Entre 30% e 55% dos pacientes nos estágios iniciais de câncer de pulmão que são submetidos à ressecção do tumor morrem pela reincidência da doença. A boa notícia é que a sobrevida desses pacientes pode aumentar significativamente com quimioterapia *adjuvante*, segundo estudos recentes [1].

Diante da difícil estimativa de sobrevida nos estágios iniciais da doença e da urgente necessidade de selecionar efetivamente pacientes que possam ser beneficiados pela quimioterapia adjuvante, métodos de prognóstico vêm sido estudados, utilizando análises sofisticadas de dados de expressão gênica que, combinadas às variáveis clínicas, podem melhorar a estimativa de sobrevida dos pacientes, segundo um grande estudo realizado nos EUA [1].

O prognóstico em câncer de pulmão é apenas uma das numerosas aplicações de dados de expressão gênica no estudo do câncer. Com esses dados, advindos de microarranjos de DNA, pode-se inferir a estrutura das redes regulatórias das interações entre DNA, RNA e proteínas, entre outras características biológicas chaves do sistema celular, cujo entendimento é essencial para o avanço do conhecimento sobre o câncer, dos métodos de diagnóstico e do desenvolvimento de novos medicamentos [2].

Com o avanço da tecnologia de microarranjos de DNA, estudos de muitos outros fenômenos biológicos têm sido realizados por meio da análise de dados de expressão gênica, domínio no qual o estudo de dependência estatística é central.

Tradicionalmente, para detectar dependência entre variáveis aleatórias, pesquisadores utilizam medidas de correlação que supõem linearidade[1]. Contudo, alguns estudos indicam que pode haver dependência não linear entre dados de expressão gênica [3], e, portanto, o uso dessas medidas de correlação poderia empobrecer a análise estatística. Não é difícil perceber que a escolha da medida de dependência a ser utilizada num experimento pode afetar significativamente os resultados.

Não só na Biologia Molecular, mas também em diversas outras áreas do conhecimento, é essencial entender as associações entre as variáveis aleatórias. Faz-se necessário, para validar uma análise num dado conjunto de dados, conhecer as características e limitações da medida de dependência utilizada. Além disso, é preciso considerar que outras medidas

[1]Quanto maior o grau de dependência linear, mais semelhante o gráfico da relação será do gráfico de uma reta. Uma associação linear perfeita é uma relação que pode ser descrita por uma equação de reta.

existentes, não necessariamente as mais tradicionais no domínio estudado, podem ser mais poderosas.

Alguns trabalhos recentes vêm ilustrando as principais diferenças de desempenho de algumas das medidas existentes. Em [3], são realizadas simulações comparando os coeficientes de correlação de Pearson [4], Spearman [5] e a medida D de Hoeffding [6], revelando que a primeira detecta relações lineares; a segunda detecta relações monotônicas [2] lineares e não lineares; e a última detecta não só relações monotônicas, mas também não monotônicas.

Em artigo publicado na *Science*, em 2011, é apresentada uma nova medida de dependência, o coeficiente de informação máxima (CIM) [7], e são feitas comparações com outras medidas, como os coeficientes de correlação de Pearson e Spearman e a informação mútua (IM) [8]. Em resposta à publicação[3], foram conduzidas simulações comparando o CIM com outros métodos, algumas realizadas por Noah Simon e Rob Tibshirani [9], com a correlação de distância (Dcor)[10] e a medida de Pearson [11]; e outras realizadas por Malka Gorfine, Ruth Heller e Yair Heller, com a medida por eles implementada (HHG) [12] e o Dcor. A conclusão desses experimentos foi que o CIM, em certos aspectos, não se mostrou mais vantajoso que o Dcor e o HHG.

Diante desses diferentes comportamentos das medidas de dependência existentes, todo pesquisador deveria conhecer as limitações e características de cada método, antes de adotar uma medida em seu trabalho. Um fator dificultante para o pesquisador é a carência de estudos abrangentes que comparem medidas tradicionais com outras recentes e que revelem o "perfil" de cada método.

A fim de suprir uma parte dessa deficiência, o trabalho proposto estabeleceu os objetivos descritos na próxima seção.

1.1 Objetivos

Neste trabalho, apresentamos um estudo comparativo entre medidas tradicionais, como as medidas de Pearson [4], Spearman [5] e Kendall [13], e outras como a correlação de distância (Dcor) [10], a informação mútua (IM) [8], o coeficiente de informação máxima (CIM) [7], e as medidas D de Hoeffding [6] e de Heller, Heller e Gorfine (HHG) [12], a fim de ilustrar as potencialidades e limitações de cada uma, considerando diversos tipos de relação (linear, monotônica não linear, não monotônica e não funcional [4]), diferentes

[2]Uma associação é monotônica se uma das variáveis aumenta ou diminui sistematicamente (isto é, apenas aumenta ou apenas diminui) quando a outra cresce.

[3]Os comentários sobre a publicação da *Science* estão disponíveis em `http://comments.sciencemag.org/content/10.1126/science.1205438`

[4]Uma associação é funcional se para cada valor da variável X, existe um único valor da variável Y associado.

tamanhos de amostra e a presença/ausência de *outliers*[5].

Além disso, ilustramos o uso das medidas utilizadas em dados de expressão gênica advindos da tecnologia de microarranjos de DNA, a partir de amostras de adenocarcinomas de pulmão.

1.2 Estrutura do presente trabalho

- No capítulo 2, fazemos uma abordagem teórica inicial para o entendimento do trabalho, com breves explicações sobre o conceito de dependência estatística, métodos estatísticos utilizados e fundamentos biológicos.
- No capítulo 3, descrevemos as medidas de dependência, os dados biológicos e a metodologia utilizados no trabalho.
- No capítulo 4 apresentamos e discutimos os resultados obtidos.
- No capítulo 5 fazemos considerações finais sobre o trabalho e sintetizamos os principais resultados.
- Por fim, a parte II trata da relação do desenvolvimento do presente trabalho com o Bacharelado em Ciência da Computação e das impressões e experiências pessoais adquiridas.

[5]Um *outlier* é uma observação numericamente distante das demais na amostra.

Capítulo 2
Preliminares

2.1 Conceitos estatísticos básicos

2.1.1 Dependência estatística entre variáveis aleatórias

Suponha que temos dados de 15 alunos de uma turma de Programação Linear do Instituto de Matemática e Estatística da USP, a partir dos quais construímos a seguinte tabela:

Tabela 2.1: Dados de 15 alunos (fictícios) de Programação Linear do IME-USP

Aluno	Média final na disciplina de Programação Linear	Altura (em cm)	Número médio de horas que o aluno passa no instituto por dia
A	0.5	175	1
B	1.2	172	8
C	2.5	181	2
D	3	162	2.5
E	3.2	183	2
F	3.2	173	3
G	3.5	178	2.5
H	4	175	3.5
I	4	166	2.5
J	4.5	174	3.3
K	5	172	4
L	5.1	177	4.5
M	8	180	12
N	9	159	8
O	10	173	10

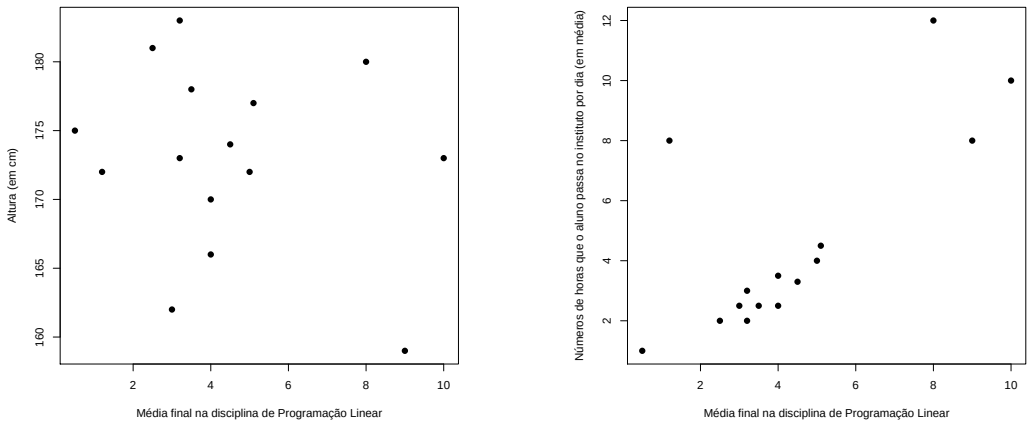
Suponha, agora, que queremos responder às seguintes perguntas:

1. Existe relação entre o desempenho na disciplina de Programação Linear e a altura do aluno?

2. Existe relação entre o desempenho na disciplina de Programação Linear e o número médio de horas que o aluno passa no instituto por dia?

Em outras palavras, queremos saber se as variáveis no nosso conjunto de dados são dependentes.

Os gráficos de dispersão na figura 2.1 sugerem que o desempenho do aluno está relacionado ao número médio de horas passadas no instituto diariamente e independe de sua altura.

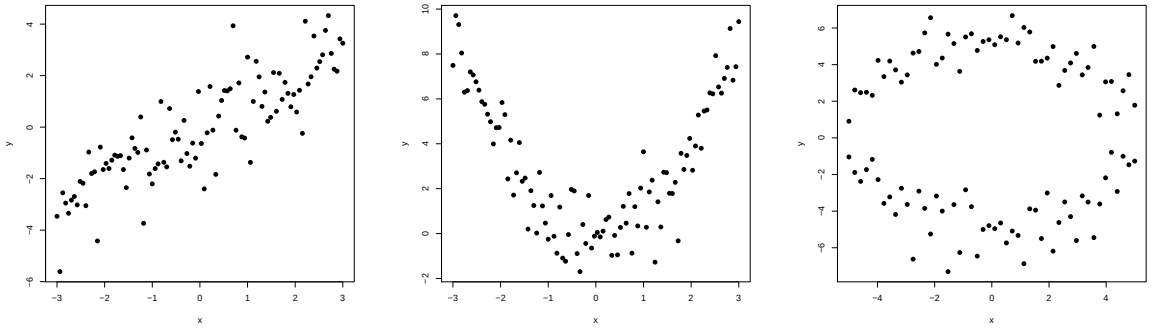


(a) Média na disciplina *versus* altura (em cm)

(b) Média na disciplina *versus* número de horas no instituto

Figura 2.1: Gráficos de dispersão a partir dos dados de 15 alunos (fictícios) de Programação Linar

De modo geral, quando duas variáveis em um conjunto de dados são dependentes, espera-se encontrar uma associação entre os valores observados, como ilustra a figura 2.2.



(a) Linear

(b) Quadrática

(c) Circunferência

Figura 2.2: Exemplo de tipos de associações entre dados

Já, quando as variáveis são independentes, não esperamos encontrar associação entre os valores observados, conforme ilustra a figura 2.3.

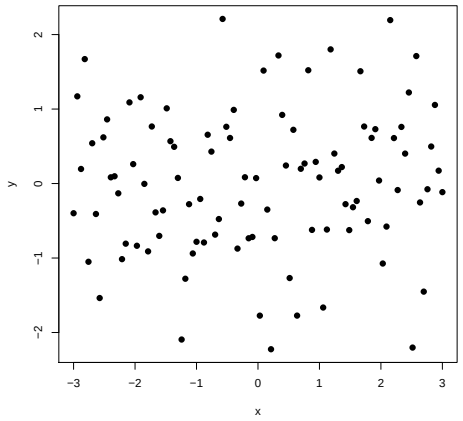


Figura 2.3: Exemplo de um conjunto de dados com variáveis independentes

Formalmente, duas variáveis aleatórias contínuas, X e Y, com funções de densidade de probabilidade $f_X(x)$ e $f_Y(y)$, respectivamente, são **independentes** se:

$$f_{XY}(x, y) = f_X(x) f_Y(y)$$

onde $f_{XY}(x, y)$ é a função de densidade de probabilidade conjunta de X e Y.

Quando $f_{XY}(x, y) \neq f_X(x) f_Y(y)$, dizemos que X e Y são **dependentes**.

2.1.2 Teste de hipóteses

Em estatística, um resultado é significante quando é improvável que ele tenha ocorrido por acaso, de acordo com um limiar de probabilidade pré-estabelecido, o nível de significância α.

Um **teste de hipóteses** é um método utilizado para avaliar se o resultado de um experimento é estatisticamente significante, isto é, se o resultado nos permitirá rejeitar a chamada hipótese nula, com uma probabilidade de erro controlada pelo valor α. O procedimento do teste se baseia nas etapas:

1. Escolhemos a **hipótese nula** (H_0), que por ora vamos supor verdadeira.

2. Explicitamos a **hipótese alternativa** (H_1), que é a hipótese que queremos sustentar.

3. Fixamos o **nível de significância do teste** α, definido como a probabilidade de cometer o **erro tipo I**, isto é, de rejeitarmos H_0, dado que H_0 é verdadeira, $\alpha = \text{P(erro tipo I)} = \text{P(rejeitar } H_0 \mid H_0 \text{ é verdadeira)}$.

4. Obtemos o **p-valor** do teste, que é a probabilidade de que o valor da estatística de teste seja pelo menos tão extremo quanto o valor observado na amostra, supondo que H_0 é verdadeira.

5. Rejeitamos a hipótese nula, caso o p-valor obtido seja menor do que o α escolhido.

Definimos $\beta = \text{P(erro tipo II)} = \text{P(não rejeitar } H_0 \mid H_0 \text{ é falsa)}$, onde o **erro tipo II** é aceitar a hipótese nula quando ela não é verdadeira.

O **poder estatístico** do teste $(1-\beta)$ é a probabilidade de rejeitarmos corretamente a hipótese nula. Esse conceito estatístico pode ser utilizado para comparar diferentes testes para uma mesma hipótese.

2.2 Técnicas computacionais e métodos estatísticos utilizados no trabalho

2.2.1 *Bootstrap* em testes de independência

Sejam X e Y duas variáveis aleatórias e $\mathbf{x}$ e $\mathbf{y}$ vetores correspondendo às suas respectivas amostras.

Considere o teste estatístico com a seguinte descrição:

$H_0 : X$ e Y são independentes

$H_1 : X$ e Y não são independentes

Suponha que a estatística r utilizada no teste de independência tenha distribuição de probabilidade desconhecida sob H_0.

Para estimarmos o p-valor do teste, podemos realizar o seguinte procedimento (*bootstrap* [14]):

1. Calulamos o valor r_0 da estatística r, a partir de $\mathbf{x}$ e $\mathbf{y}$.

2. Fixamos um número B de iterações (permutações).

3. A cada iteração i, realizamos uma permutação aleatória de $\mathbf{y}$ e obtemos um novo vetor $\mathbf{y^*}$. Calculamos r_i a partir de $\mathbf{x}$ e $\mathbf{y^*}$.

4. Ao final das B iterações, calculamos o p-valor como sendo a proporção dos valores de r_i maiores que r_0.

Nas simulações realizadas nesse trabalho, utilizamos $B = 1000$.

2.2.2 Múltiplos testes

Quando múltiplos testes de hipóteses são realizados em um experimento, é preciso considerar a probabilidade de rejeitar incorretamente alguma hipótese nula em todo o conjunto de testes.

Suponha que fixamos $\alpha = 0,05$. Então, em 100 testes independentes entre si, onde as hipóteses nulas são verdadeiras, a probabilidade de não cometermos o erro tipo I é $(1 - 0,05)^{100} = 0,006$. Desse modo, a probabilidade de rejeitar incorretamente pelo menos uma hipótese nula é $1 - 0,006 = 0,994$, um valor muito superior ao α desejado.

Um método utilizado para controlar a proporção de ocorrências do erro tipo I, ou a **taxa de falsos-positivos** em múltiplos testes de hipóteses independentes, é o método FDR de Benjamini–Hochberg [15] descrito a seguir.

1. Considere as hipóteses sendo testadas H_1, H_2, ..., H_m e os respectivos p-valores P_1, P_2, ..., P_m.
2. Sejam $P_{(1)} \leq P_{(2)} \leq ... \leq P_{(m)}$ os p-valores ordenados.
3. Seja k o maior i tal que $P_{(i)} \leq \frac{i}{m}\alpha$
4. Rejeitamos $H_{(i)}$, para $i = 1, 2, ..., k$.

Equivalentemente, podemos "ajustar" os p-valores dos testes, com o seguinte procedimento:

$$\begin{aligned} \tilde{P}_{(m)} &= P_{(m)} \\ \tilde{P}_{(m-1)} &= min(\tilde{P}_{(m)}, \tfrac{m}{m-1}P_{(m-1)}) \\ &\vdots \\ \tilde{P}_{(1)} &= min(\tilde{P}_{(2)}, mP_{(1)}) \end{aligned}$$

2.2.3 Curva ROC

A curva ROC (*Receiver Operating Characteristics*) [16] é utilizada para avaliar o desempenho de classificadores binários.

Dados uma instância e um classificador que rotula em *positivo* ou *negativo*, existem quatro configurações possíveis. Se a instância for corretamente classificada como *positivo*, ela é um **verdadeiro positivo**. Se for incorretamente classificada como *positivo*, ela é um **falso positivo**. Caso a instância seja classificada como *negativo* e de fato pertença à classe *negativo*, ela é um **verdadeiro negativo**. Se a instância for incorretamente classificada como *negativo*, ela é um **falso negativo**. As quatro configurações aqui expostas podem ser resumidas na tabela de contingência 2.2 exibida a seguir.

Tabela 2.2: Tabela de contingência

	Valor Verdadeiro	
Resultado do teste	positivo	negativo
positivo	VP Verdadeiro Positivo	FP Falso Positivo
negativo	FN Falso Negativo	VN Verdadeiro Negativo

Definimos:

- **Taxa de verdadeiro positivo**: $TVP = \frac{VP}{\text{Total de positivos}} = \frac{VP}{VP+FN}$
- **Taxa de falso positivo**: $TFP = \frac{FP}{\text{Total de negativos}} = \frac{FP}{FP+VN}$

A **curva ROC** é desenhada na grade da taxa de falso positivo *versus* a taxa de verdadeiro positivo. Cada ponto da curva está associado a um limiar para a classificação em *positivo* e *negativo* (supondo que o resultado devolvido pelo classificador é um valor utilizado para discriminar as duas classes).

Uma curva ROC que passa pela coordenada $(0, 1)$ representa o mellhor desempenho possível de um classificador, pois indica uma taxa de verdadeiro positivo de 100% (isto é, ausência de falsos negativos), e uma taxa de falso positivo de 0%.

Um classificador que rotula a instância em *positivo* ou *negativo* aleatoriamente, teria uma curva ROC na linha diagonal, com os extremos nas coordenadas $(0, 0)$ e $(1, 0)$, como indica a linha tracejada na figura 2.4.

Assim, um classificador com desempenho melhor do que uma rotulação aleatória, deve ter uma curva ROC correspondente acima da linha diagonal. O desenho da curva nos permite, então, comparar diferentes classificadores. A figura 2.4, ilustra curvas ROC construídas com dois classificadores: o classificador 1, que apresenta bom desempenho (a curva correspondente está bem acima da diagonal) e o classificador 2, que é ineficaz (sua curva está próxima da linha diagonal).

Sintetizar a informação desse gráfico em um valor escalar pode ser interessante, especialmente quando muitas curvas são analizadas. Um método comum para sintetizar esse resultado é a **área sob a curva ROC**, AUC (*Area Under Curve*). Como a AUC é uma porção de um quadrado unitário, seu valor está entre 0 e 1. No caso da rotulação aleatória espera-se que o valor da AUC seja 0,5 e, quanto mais próximo o valor da AUC estiver de 1, melhor o desempenho do classificador. A área sob a curva pode também ser

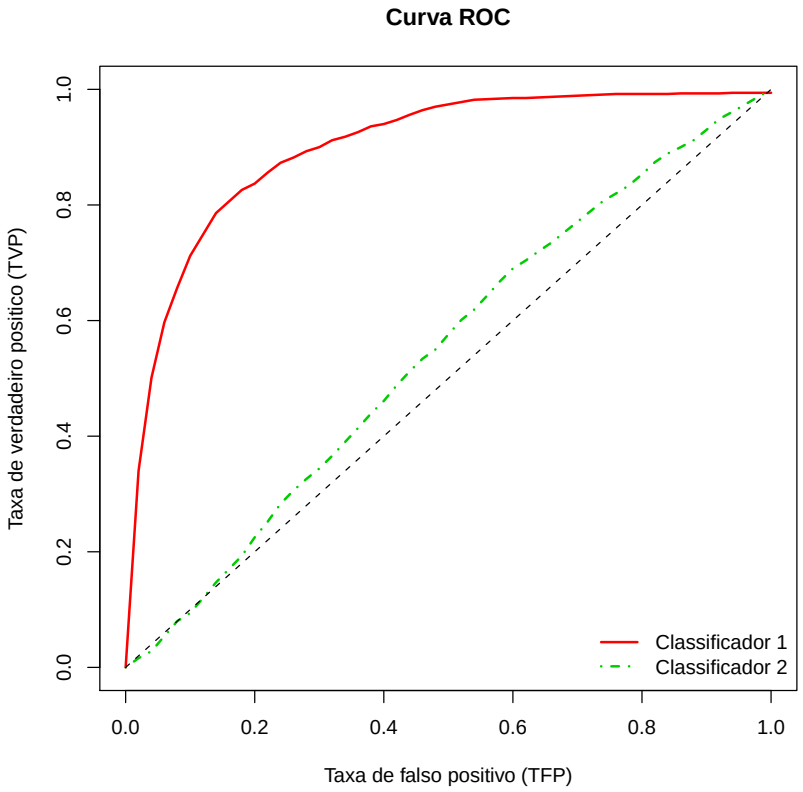


Figura 2.4: Curvas ROC construídas a partir de dois classificadores

interpretada como a probabilidade de que um classificador dê uma "pontuação" maior para instâncias positivas do que para instâncias negativas.

2.3 Fundamentos biológicos

Uma **célula** representa a menor unidade de vida. Nessa estrutura, estão contidas as características morfológicas e fisiológicas dos organismos vivos. Assim, as propriedades de um dado organismo dependem de suas células individuais, cuja continuidade ocorre através de seu material genético.

2.3.1 Ácidos nucléicos

Os **ácidos nucléicos** são moléculas responsáveis por estocar e transmitir a informação genética na célula. A partir dessas moléculas, as células são instruídas sobre quais proteínas sintetizar e em que quantidade. Existem dois tipos de ácidos nucléicos: o ácido desoxirribonucléico (DNA) e o ácido ribonucléico (RNA). O último é basicamente dividido nas classes RNA mensageiro (mRNA), RNA transportador (tRNA) e RNA ribossômico

(rRNA).

O **gene** corresponde a uma sequência particular de DNA, codificadora de uma informação, proteína ou RNA.

2.3.2 Expressão gênica

Expressão gênica é o processo pelo qual a informação de um gene é utilizada na síntese de um produto gênico funcional, como as proteínas.

A expressão gênica envolve a cópia de regiões do DNA (genes) em uma molécula de mRNA e a passagem da informação existente nesse mRNA para uma sequência de aminoácidos (proteína). A síntese das moléculas de RNA ocorre por um processo chamado **transcrição** e a síntese de proteínas é feita pelo processo de **tradução**. A figura 2.5 esquematiza o fluxo de informação genética.

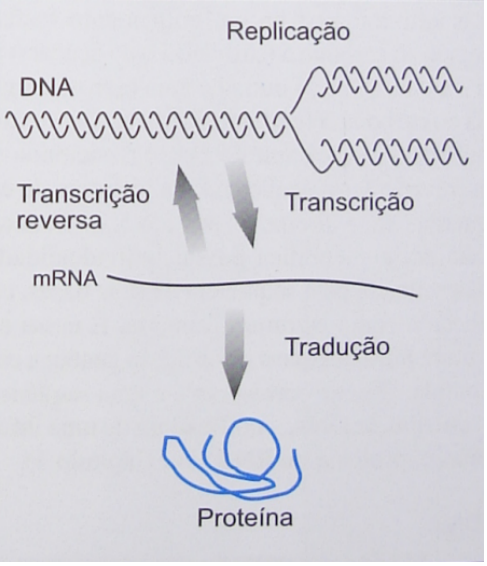


Figura 2.5: Esquema de fluxo de informação genética. O DNA pode ser replicado ou transcrito em RNA. Os mRNA são traduzidos em sequências protéicas. Em algumas situações especiais, o RNA pode ser reversamente transcrito produzindo DNA. Figura retirada de [17], página 34

O seguinte trecho retirado das páginas 34 e 35, em [17], elucida os efeitos da expressão de um gene:

> A quantidade de mRNA produzido a partir de uma região particular do DNA é controlada por proteínas regulatórias, que se ligam a sítios específicos no DNA. Em qualquer célula, em qualquer tempo, alguns genes são utilizados para transcrever RNA em grandes quantidades, enquanto outros não são tão ativamente, ou mesmo não são transcritos. A partir de um gene ativo, milhares de mRNAs podem ser sintetizados em cada cécula. Como cada molécula de RNA pode ser traduzida em milhares de cópias de uma cadeia polipeptídica, a informação

contida em uma pequena região do DNA pode dirigir a síntese de milhões de cópias de uma proteína específica.

2.3.3 Microarranjos de DNA

Um microarranjo de DNA é um arranjo pré-definido de moléculas de DNA ligadas à uma lâmina. Esse arranjo é construído para medir os níveis de expressão de vários genes simultaneamente.

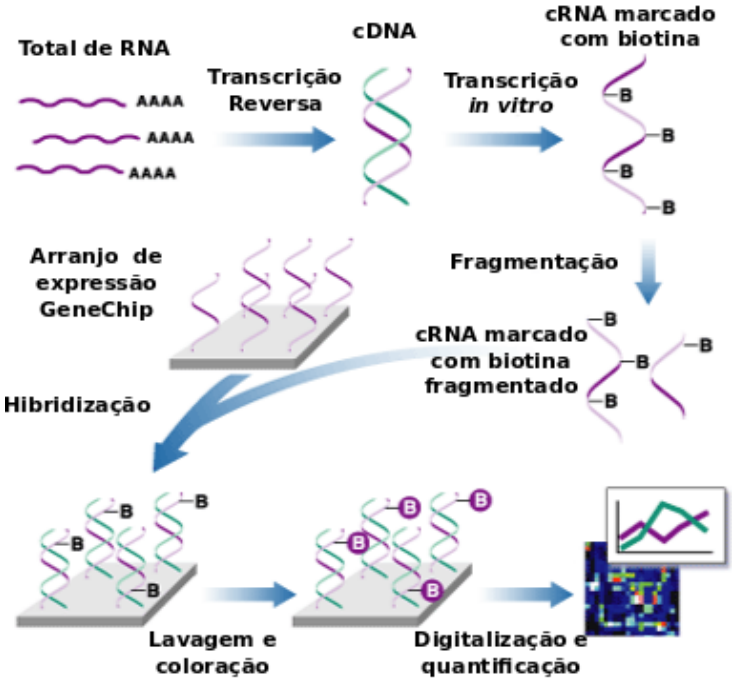


Figura 2.6: Esquema ilustrativo da técnica de microarranjo de DNA na plataforma *Affymetrix*. Retirado e modificado de `http://angerer.swissbrain.org/expression_oveview.gif`

O experimento (esquematizado na figura 2.6) segue os seguintes passos:

1. O RNA é extraído de uma amostra biológica.
2. A partir de um processo de transcrição reversa, é sintetizado o DNA complementar (cDNA)
3. O cDNA produzido é transcrito *in vitro* para cRNA marcado com biotina.
4. O cRNA produzido é fragmentado e hibridizado.
5. A parte não hibridizada é removida do arranjo.
6. O arranjo passa por processos de lavagem e coloração

7. É gerada uma imagem a partir do microarranjo de DNA produzido

Por fim, a partir da imagem gerada, obtemos uma quantificação do mRNA produzido por cada gene.

Capítulo 3
Medidas de dependência

O estudo comparativo se baseia na avaliação do poder estatístico de medidas de dependência[1] em diversos tipos de dados simulados com a ferramenta R [18]. Tais dados simulam amostras de duas variáveis aleatórias X e Y, que são submetidas ao teste estatístico com a seguinte descrição:

$H_0 : X$ e Y são independentes

$H_1 : X$ e Y não são independentes

Os testes de independência são realizados em diversas situações simuladas, considerando-se diferentes tamanhos de amostra, presença/ausência de *outliers*, e diferentes tipos de associação (linear, monotônica, não monotônica, não funcional, ou correlação local) a fim de caracterizar o desempenho de cada medida.

A medida de desempenho se baseia no poder estatístico dos métodos estudados, que é estimado pela proporção de vezes que a hipótese nula é rejeitada em 1000 simulações, dado o nível de significância α.

Variando o α de 0 a 1, construimos uma curva do poder estimado a partir dos 1000 testes de independência. Podemos interpretar a curva do poder como uma curva ROC, tomando o poder como a taxa de verdadeiros positivos e o nível de significância como simultaneamente a taxa de falsos positivos e o limiar do p-valor para a classificação das variáveis X e Y em dependentes e não dependentes.

Como medida de comparação entre os métodos, adotamos a área sob a curva (AUC) produzida, que nos fornece uma medida geral do desempenho de cada método na situação em que foi gerada a curva. Quanto maior a área obtida, maior o poder estatístico.

A seguir, explicaremos os testes realizados para cada método. Considere $\mathbf{x}$ e $\mathbf{y}$ vetores de tamanho n que correspondem às amostras de X e Y, respectivamente. Denotaremos (x_i, y_i) para os pares de valores observados nas amostras.

[1] Chamaremos de poder de uma medida de dependência, o poder estatístico do teste de independência realizado com essa medida.

3.1 Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida de correlação entre duas variáveis aleatórias, utilizada para medir o quanto uma associação pode ser descrita como uma função linear ou, em outras palavras, o quão forte é a dependência linear entre as variáveis. A referida medida é definida como a razão entre a covariância das duas variáveis e o produto de seus respectivos desvios padrão.

A correlação de Pearson aplicada nas amostras $\mathbf{x}$ e $\mathbf{y}$ é dada por:

$$r_p = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2}\sqrt{\sum (y_i - \bar{y})^2}} = \frac{n\sum x_i y_i - \sum x_i \sum y_i}{\sqrt{n\sum x_i^2 - (\sum x_i)^2}\sqrt{n\sum y_i^2 - (\sum y_i)^2}}$$

O valor do coeficiente está entre -1 e 1. Quando r_p vale -1, todos os pontos (x_i, y_i) estão em uma reta e os valores de $\mathbf{y}$ diminuem à medida que os valores de $\mathbf{x}$ aumentam. Se r_p vale 0, não há correlação linear. Quando r_p vale 1, os valores de $\mathbf{y}$ aumentam quando os valores de $\mathbf{x}$ aumentam e há uma reta que passa por todos os pontos (x_i, y_i). Valores intermediários podem indicar o quão forte é a associação linear entre $\mathbf{x}$ e $\mathbf{y}$.

Para o teste de independência, definimos:

$$t = \frac{r_p\sqrt{n-2}}{\sqrt{1 - r_p^2}}$$

Sob H_0, isto é, se X e Y são intependentes, t segue uma distribuição t de Student com $n - 2$ graus de liberdade.

A partir dessa distribuição obtemos o p-valor associado ao teste.

3.2 Correlação de distância (Dcor)

A correlação de distância é uma medida de dependência entre duas variáveis aleatórias análoga à correlação de Pearson.

A correlação de distância aplicada às amostras $\mathbf{x}$ e $\mathbf{y}$ é dada por:

$$dCor(\mathbf{x}, \mathbf{y}) = \frac{dCov(\mathbf{x}, \mathbf{y})}{\sqrt{dVar(\mathbf{x})dVar(\mathbf{y})}}$$

onde:

- $dCov_n^2(\mathbf{x}, \mathbf{y}) := \frac{1}{n^2}\sum_{k,l} A_{k,l}B_{k,l}$
- $dVar_n^2(\mathbf{x}) := dCov_n^2(\mathbf{x}, \mathbf{x})$

- $A_{k,l} = a_{k,l} - \bar{a}_{k.} - \bar{a}_{.l} + \bar{a}_{..}$
- $a_{k,l} = \|x_k - x_l\|$, para $k, l = 1, 2, ..., n$
- $B_{k,l} = b_{k,l} - \bar{b}_{k.} - \bar{b}_{.l} + \bar{b}_{..}$
- $b_{k,l} = \|y_k - y_l\|$, para $k, l = 1, 2, ..., n$
- $\|.\|$ é a norma Euclidiana
- $\bar{a}_{k.}$ e $\bar{b}_{k.}$ são os valores médios da k-ésima linha das matrizes de distância $(a_{k,l})$ e $(b_{k,l})$, respectivamente
- $\bar{a}_{.l}$ e $\bar{b}_{.l}$ são os valores médios da k-ésima coluna das respectivas matrizes de distância
- $\bar{a}_{..}$ e $\bar{b}_{..}$ são os valores médios das matrizes $(a_{k,l})$ e $(b_{k,l})$, respectivamente

$nDcov^2$ é uma forma quadrática de variáveis aleatórias gaussianas, com coeficientes que dependem da distribuição de X e Y. Quando as distribuições são desconhecidas o teste de independência baseado em $nDcov^2$ pode ser produzido com a técnica de *bootstrap*.

Nos nossos experimentos, utilizamos a implementação do teste estatístico do pacote *energy* [19] do R.

3.3 Correlação de Spearman

O coeficiente de correlação de Spearman é uma medida de dependência estatística entre duas variáveis que pode indicar o quão bem uma relação pode ser descrita como uma função monotônica.

A correlação de Spearman, denotada por r_s, é a aplicação do coeficiente de correlação de Pearson nos dados convertidos em postos [2].

Se não há valores repetidos nas amostras, pode-se obter o coeficiente de Spearman com a seguinte fórmula:

$$r_s = 1 - 6\frac{\sum d_i^2}{n(n^2 - 1)}$$

onde d_i é a diferença entre os postos dos valores correspondentes de $\mathbf{x}$ e $\mathbf{y}$.

O coeficiente de correlação de Spearman verifica, portanto, se há dependência linear entre os postos dos valores observados, o que corresponde a verificar se os valores de $\mathbf{y}$ crescem quando $\mathbf{x}$ cresce, ou se os valores de $\mathbf{y}$ diminuem, à medida que os valores de $\mathbf{x}$ aumentam.

[2] O posto de um elemento de um vetor é a posição do mesmo no vetor com os dados em ordem crescente.

Como são utilizados os postos no lugar dos dados, a presença de *outliers* não deve provocar grandes alterações no valor de r_s.

Para o teste de independência, definimos:

$$t = \frac{r_s\sqrt{n-2}}{\sqrt{1-r_s^2}}$$

Sob H_0, t segue uma distribuição t de Student com $n-2$ graus de liberdade.

A partir dessa distribuição obtemos o p-valor associado ao teste.

3.4 Tau de Kendall

O coeficiente de Kendall é uma estatística usada para medir a associação entre duas variáveis aleatórias. Assim como o coeficiente de correlação de Spearman, a medida de Kendall se baseia nos postos dos dados. Contudo, veremos que a interpretação dessas medidas são diferentes.

Obtemos o tau de Kendall a partir da seguinte fórmula:

$$\tau = \frac{C-D}{N}$$

onde C é o número de pares concordantes, D é o número de pares discordantes e N é o número total de pares.

Dois pares de observação quaisquer (x_i, y_i) e (x_j, y_j) são concordantes se $x_i > x_j$ e $y_i > y_j$ ou $x_i < x_j$ e $y_i < y_j$; e discordantes se $x_i > x_j$ e $y_i < y_j$ ou $x_i < x_j$ e $y_i > y_j$. Se não há valores repetidos nas amostras:

$$N = \frac{1}{2}n(n-1)$$

O valor de τ está entre -1 e 1. Se os postos de $\mathbf{x}$ são os mesmos de $\mathbf{y}$, então τ vale 1. Se os postos de $\mathbf{x}$ são o reverso dos postos de $\mathbf{y}$, então a medida de Kendall vale -1. Se X e Y são independentes, o valor esperado do coeficiente aplicado às amostras de X e Y é zero.

A distribuição de τ, sob H_0, para n suficientemente grande, pode ser aproximada para uma normal com média 0 e variância $\frac{2(2n+5)}{9n(n-1)}$.

Para amostras pequenas, a distribuição de τ pode ser obtida explicitando-se todas as $n!$ possíveis permutações dos postos das observações.

Dessa forma, obtemos o p-valor associado ao teste de independência.

Nos experimentos desse trabalho foi utilizada a implementação do pacote *stats* do R.

3.5 Medida D de Hoeffding

A medida D de Hoeffding é uma medida de associação entre duas variáveis aleatórias, que é calculada a partir das amostras x e y, com a seguinte fórmula:

$$D = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)}$$

onde:

- $D_1 = \sum_{i=1}^{n} Q_i(Q_i - 1)$
- $D_2 = \sum_{i=1}^{n} (R_i - 1)(R_i - 2)(S_i - 1)(S_i - 2)$
- $D_3 = \sum_{i=1}^{n} (R_i - 2)(S_i - 2)Q_i$
- R_i é o posto de x_i
- S_i é o posto de y_i
- Q_i é o número de pontos com ambos os valores de x e y menores do que o i-ésimo ponto.

Sejam $F_{XY}(x, y)$ a função de distribuição acumulada conjunta de X e Y e $F_X(x)$ e $F_Y(y)$ as funções de distribuição acumulada marginais de X e Y, respectivamente. D é uma medida da distância entre $F_{XY}(x, y)$ e $F_X(x)F_Y(y)$.

Sob H_0, ou seja, sob a hipótese de que X e Y são independentes:

$$F_{XY}(x, y) = F_X(x)F_Y(y)$$

Seja ρ_n o menor valor satisfazendo a desigualdade:

$$P\{D > \rho_n | F_{XY}(x, y) = F_X(x)F_Y(y)\}$$

onde P é a distribuição de probabilidade de D.

Tal valor satisfaz:

$$30\rho_n \leq \sqrt{\frac{2(n^2 + 5n - 32)}{9n(n-1)(n-3)(n-4)\alpha}}$$

onde α é o nível de significância do teste

Rejeitamos H_0 se, e somente se, $D > \rho_n$.

Obtemos, assim, um teste de independência consistente [3], segundo Hoeffding.

Neste trabalho, utilizamos implementação do pacote *Hmisc* [20] do R.

[3] O teste de uma hipótese H_0 é consistente se a probabilidade de aceitar H_0, quando a hipótese alternativa é verdadeira, tende a zero à medida que o tamanho da amostra aumenta.

3.6 Medida de Heller, Heller e Gorfine (HHG)

Heller, Heller e Gorfine propõem um teste de independência consistente e com elevado poder estatístico.

O teste se baseia nas distâncias entre os valores de $\mathbf{x}$ e os valores de $\mathbf{y}$, respectivamente, $\{d_x(x_i, x_j) : i, j \in \{1, ..., n\}\}$ e $\{d_y(y_i, y_j) : i, j \in \{1, ..., n\}\}$.

Para cada observação i e cada $j \neq i$, $1 \leq j \leq n$, definimos $R_x(i, j) = d_x(x_i, x_j)$ e $R_y(i, j) = d_y(y_i, y_j)$, $A_{11}(i, j) = \sum_{k=1, k\neq i, k\neq j}^{n} I\{d(x_i, x_k) \leq d(x_i, x_j)\} I\{d(y_i, y_k) \leq d(y_i, y_j)\}$, A_{12}, A_{21} e A_{22} são definidas de modo análogo, e $A_{m.}$ e $A_{.m}$, $m = 1, 2$, são as somas da linha m e coluna m, respectivamente.

- $I\{d(x_i, x_k) \leq d(x_i, x_j)\}$ vale 1, se $d(x_i, x_k) \leq d(x_i, x_j)$ e vale zero, caso contrário
- $I\{d(y_i, y_k) \leq d(y_i, x_j)\}$ vale 1, se $d(y_i, y_k) \leq d(y_i, x_j)$ e vale zero, caso contrário

Seja

$$S(i, j) = \frac{(n-2)\{A_{12}(i, j)A_{21}(i, j) - A_{11}(i, j)A_{22}(i, j)\}^2}{A_{1.}(i, j)A_{2.}(i, j)A_{.1}(i, j)A_{.2}(i, j)}$$

Para estimar o p-valor, sob H_0, pode-se utilizar o método de *bootstrap* com a seguinte estatística:

$$T = \sum_{i=1}^{n} \sum_{j=1}^{n} S(i, j)$$

As simulações realizadas nesse trabalho utilizaram o pacote *HHG2x2* que pode ser solicitado aos autores.

3.7 Informação mútua (IM)

A informação mútua de duas variáveis aleatórias é uma medida da dependência mútua entre as mesmas.

Se X e Y são variáveis contínuas, a informação mútua é dada por:

$$I(X, Y) = \int_Y \int_X f_X Y(x, y) \log \frac{f_{XY}(x, y)}{f_X(x) f_Y(y)} dxdy,$$

onde $f_{XY}(x, y)$ é a função de densidade de probabilidade conjunta de X e Y e $f_X(x)$ e $f_Y(y)$ são as funções de densidade de probabilidade marginais de X e Y, respectivamente.

No caso discreto, temos:

$$I(X, Y) = \sum_{y \in Y} \sum_{x \in X} p(x, y) \log \frac{p(x, y)}{p(x)p(y)} dxdy,$$

onde $p_{XY}(x, y)$ é a função de distribuição de probabilidade conjunta de X e Y e $p_X(x)$ e $p_Y(y)$ são as funções de distribuição de probabilidade marginais de X e Y, respectivamente.

A informação mútua mede o quanto o conhecimento sobre uma variável diminui a incerteza sobre a outra. Se X e Y são independentes conhecer uma delas não traz informação sobre a outra. Se X e Y são indênticas, então o conhecimento sobre uma determina a outra.

Podemos listar as seguintes propriedades de $I(X, Y)$:

1. $I(X, Y) = 0$ se, e só se, X e Y são independentes
2. $I(X, Y) \geq 0$
3. $I(X, Y) = I(Y, X)$

A partir das amostras $\mathbf{x}$ e $\mathbf{y}$ estimamos a informação mútua entre X e Y empiricamente (o que neste trabalho é feito com o pacote *entropy* [21]). Como não se conhece uma fórmula analítica da distribuição de probabilidade dessa medida, construímos o teste de independência com base na técnica computacional de *bootstrap*.

3.8 Coeficiente de Informação Máxima (CIM)

O Coeficiente de Informação Máxima (CIM) é uma medida de associação entre duas variáveis aleatórias que, quando aplicada em dados com diferentes tipos de associações, mas mesmo ruído, produz avaliações semelhantes.

Considere o conjunto D dos pares de observação (x_i, y_i). Uma grade a-por-b, é um par que contém uma partição de D nos valores de $\mathbf{x}$ formada por a partes ($\mathbf{x}$-partição) e uma partição de D nos valores de $\mathbf{y}$ composta por b partes ($\mathbf{y}$-partição). Chamaremos a intersecção entre uma parte da $\mathbf{x}$-partição e uma parte da $\mathbf{y}$-partição de célula.

O Coeficiente de Informação Máxima de D é dado por:

$$CIM(D) = \max_{ab<B(n)} M(D)_{a,b}$$

onde $1 < B(n) \leq n^{0.6}$ e M(D) é uma matriz cujas entadas são:

$$M(D)_{a,b} = \frac{I^*(D, a, b)}{\log \min\{a, b\}}$$

$I^*(D, a, b)$ é a maior informação mútua entre todas as grades a-por-b, tomando-se como função de distribuição de probabidade a fração dos pontos em D que estão na célula da grade em que se encontra certo ponto (x_i, y_i). O valor de CIM está entre 0 e 1.

Intuitivamente, o CIM se baseia na ideia de que se uma relação entre duas variáveis existe, então deve haver uma grade que encapsula a associação entre as mesmas.

Neste trabalho, utlilizamos uma aproximação do CIM implementada pelo programa *MINE* [22].

Como não se conhece uma fórmula fechada para a distribuição de probabilidade do CIM, o teste de independência é feito com base na técnica computacional de *bootstrap*.

3.9 Simulações

Foram realizados 1000 testes de independência com cada medida e cada uma das situações descritas mais adiante. Denotaremos $\varepsilon \sim N(\mu, \sigma)$ e $\varepsilon \sim U(a, b)$ para ε variável aleatória que segue, respectivamente, uma distribuição normal com média 0 e desvio padrão 1 e uma distribuição uniforme no intervalo $[-1, 1]$.

Simulações realizadas com amostras de tamanhos 10, 20, 30 e 50

(a) Independente:
x_i varia de -3 a 3, em intervalos iguais
$y_i = \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$

(b) Independente com *outliers*:
x_i varia de -3 a 3, em intervalos iguais
$y_i = \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$
Foram introduzidos *outliers* em 7% da amostra, nas posições finais: $y_i \sim N(0, 100)$

(c) Linear:
x_i varia de $-1,5$ a $2,5$, em intervalos iguais
$y_i = 0,5x_i + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$

(d) Linear com *outliers*:
x_i varia de -3 a 3, em intervalos iguais
$y_i = 0,5x_i + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$
Foram introduzidos *outliers* em 7% da amostra, nas posições finais: $y_i \sim N(0, 100)$

(e) Exponencial:
x_i varia de -30 a 20, em intervalos iguais
$y_i = 0,01e^{x_i} + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$

(f) Quadrática:
x_i varia de -3 a 3, em intervalos iguais

$y_i = x_i^2 + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0,1)$

(g) Quadrática com *outliers*:
x_i varia de -2 a 2, em intervalos iguais
$y_i = x_i^2 + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0,1)$
Foram introduzidos *outliers* em 7% da amostra, nas posições do "meio": $y_i \sim N(0,100)$

(h) Seno:
x_i varia de 0 a 10, em intervalos iguais
$y_i = 2seno(x_i) + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim U(-1,1)$

(i) Cúbica:
x_i varia de $0,4$ a $1,6$, em intervalos iguais
$y_i = 30(x_i - 0,5)(x_i - 1)(x_i - 1,5) + \varepsilon_i$, onde ε_i é uma observação de $\varepsilon \sim N(0,1)$

(j) Circunferência:
x_i varia de -5 a 5, e de 5 a -5, em intervalos iguais
$y_i = \sqrt{25 - x_i^2} + \varepsilon_i$, para i de 1 a $\frac{n}{2}$
$y_i = -\sqrt{25 - x_i^2} + \varepsilon_i$, para i de $\frac{n}{2} + 1$ a n, onde ε_i é uma observação de $\varepsilon \sim N(0,1)$

(k) X:
x_i varia de -5 a 5, e de 5 a -5, em intervalos iguais
$y_i = x_i + \varepsilon_i$, para i de 1 a $\frac{n}{2}$
$y_i = -x_i + \varepsilon_i$, para i de $\frac{n}{2} + 1$ a n, onde ε_i é uma observação de $\varepsilon \sim U(-1,1)$

Simulações realizadas com amostras de tamanho 40 e 140

(l) Quadrado:
x_i varia de 6 a 9, em intervalos iguais, para i de 1 a $\frac{n}{4}$
x_i vale 9, para i de $\frac{n}{4} + 1$ a $\frac{2n}{4}$
x_i varia de 9 a 6, em intervalos iguais, para i de $\frac{2n}{4} + 1$ a $\frac{3n}{4}$
x_i vale 6, para i de $\frac{3n}{4} + 1$ a n
$y_i = 6 + \varepsilon_i$, para i de 1 a $\frac{n}{4}$
$y_i = w_i + \varepsilon_i$, e w varia de 6 a 9, em intervalos iguais, para i de $\frac{n}{4} + 1$ a $\frac{2n}{4}$
$y_i = 9 + \varepsilon_i$, para i de $\frac{2n}{4} + 1$ a $\frac{3n}{4}$
$y_i = w_i + \varepsilon_i$, e w varia de 9 a 6, em intervalos iguais, para i de $\frac{3n}{4} + 1$ a n
onde ε_i é uma observação de $\varepsilon \sim U(-1,1)$

Simulações realizadas com amostras de tamanho 100

(m) Correlação local:

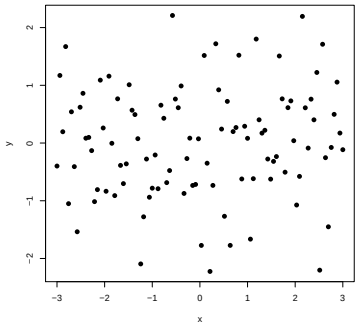
x_i varia de 3 a 6, em intervalos iguais

$y_i = \varepsilon_i$, para i de 1 a 40 e 61 a 100

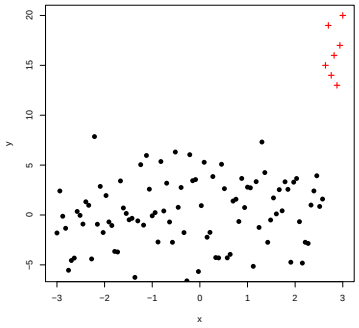
$y_i = x_i + \varepsilon_i$, para i de 41 a 60

onde ε_i é uma observação de $\varepsilon \sim N(0, 1)$

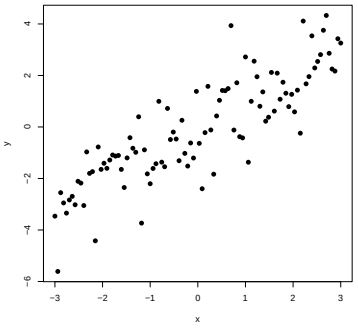
A figura 3.1 ilustra um exemplo de cada associação simulada.



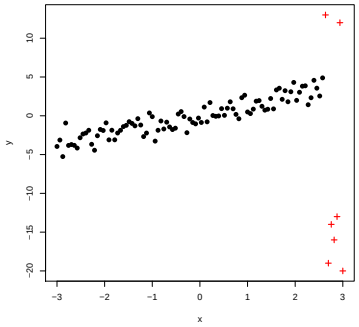
(a) Independente



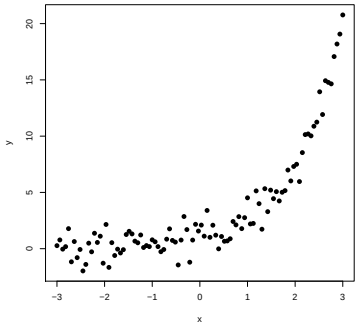
(b) Independente com *outliers*



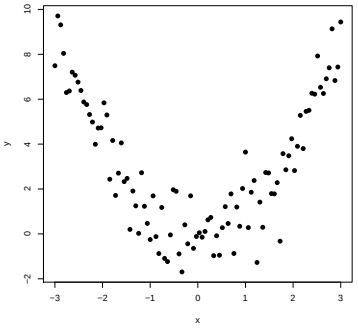
(c) Linear



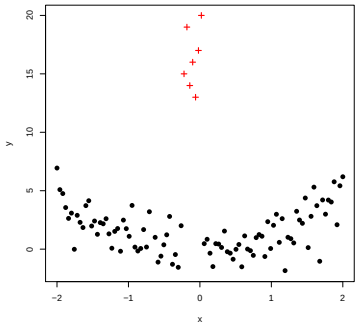
(d) Linear com *outliers*



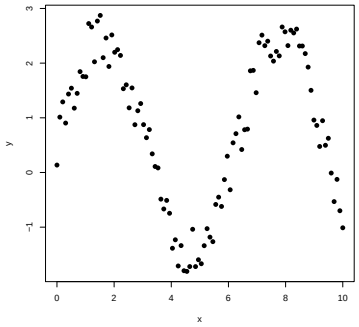
(e) Exponencial



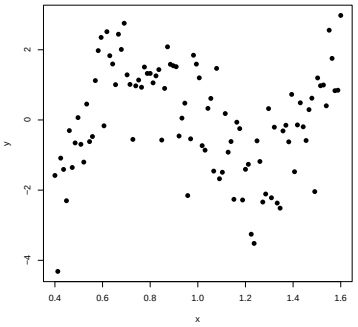
(f) Quadrática



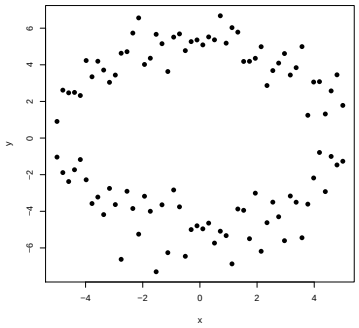
(g) Quadrática com *outliers*



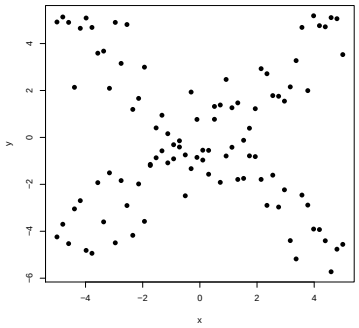
(h) Seno



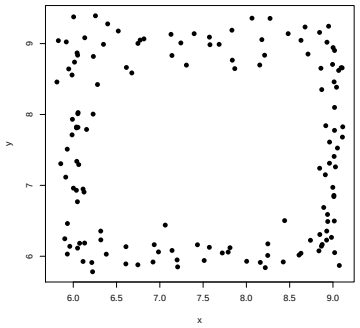
(i) Cúbica



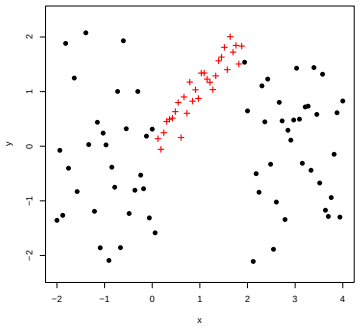
(j) Circunferência



(k) X



(l) Quadrado



(m) Correlação local

Figura 3.1: Ilustração das simulações. Os pontos em vermelho, na forma de cruz, indicam *outliers* nas figuras (b), (d) e (g), e a correlação linear na figura (m).

3.10 Dados de expressão gênica

A fim de ilustrar o uso das medidas de dependência, selecionamos dados biológicos reais, que resultam de experimentos com microarranjos de DNA a partir de 104 amostras de adenocarcinomas de pulmão (estágio I) [1], realizados nos laboratórios do *Moffitt Cancer Center* (HLM) e do *Memorial Sloan-Kettering Cancer Center* (MKCC). Os dados de expressão gênica das referidas amostras foram retirados do *caArray Data Portal* [23] e podem ser localizados pelo ID `jacob-00182`.

Nosso interesse nesse experimento é verificar associações entre o gene WNT5A, que é conhecido na literatura por estar associado ao cancer de pulmão [24], e outros 75 genes selecionados já descritos na literatura como relacionados com o WNT5A. Esta parte do trabalho teve o auxílio da pesquisadora Asuka Nakata, do Instituto de Ciências Médicas da Universidade de Tóquio.

Os dados foram processados com a função *justRMA* (para normalização e sumarização) do pacote *affy* [25] do *Bioconductor*. Utilizamos o pacote *u133agbcdf* [26] para agrupar as

"sondas"[4] do *Affymetrix* por gene. As amostras de tumor no estágio I da doença foram selecionadas de acordo com o sistema de classificação TNM [27].

Por fim, para cada medida, realizamos testes de independência com o gene WNT5A e cada um dos 75 genes selecionados.

Como os experimentos de microarranjos de DNA foram realizados em dois laboratórios diferentes (HLM e MKCC), realizamos testes separados para cada laboratório, após a remoção de *outliers*. As observações removidas foram as que apresentaram nível de expressão gênica menor do que $q_1 - \frac{3}{2}d_q$ e maiores do que $q_3 + \frac{3}{2}d_q$, onde q_1 e q_3 são, respectivamente, o primeiro e terceiro quartil, e $d_q = q_3 - q_1$ [28].

Para cada gene, temos dois p-valores correspondentes aos dois laboratórios. A fim de combinar esses dois resultados, utilizamos o método de Fisher [29] descrito a seguir:

1. Definimos: $\chi^2 = -2\sum_{i=1}^{k} log_e(p_i)$, onde p_i é o p-valor do i-ésimo teste de hipóteses e k é o número de testes (no nosso caso $k = 2$).

2. Como os testes são independentes, supondo que a hipótese nula é verdadeira (que os genes sendo testados não tem associação), χ^2 tem uma distribuição qui-quadrado com $2k$ graus de liberdade.

3. O p-valor resultante é obtido a partir da distribuição de χ^2.

Calculados os p-valores, utilizamos a implementação do método FDR do pacote *stats* do R para ajustá-los, a fim de controlar a taxa de falsos positivos dos 75 testes correspondentes aos 75 genes escolhidos.

Para verificar a consistência do experimento, realizamos também testes com genes de controle, com os quais não se espera detectar dependência.

[4] "Sondas"são fragmentos do cDNA fixado à lâmina, os quais contém, cada um, as seqüências de um único gene. Nos microarranjos, cada gene é representado por mais de uma "sonda".

Capítulo 4
Resultados e discussões

4.1 Estudo do desempenho das medidas nas simulações

Os resultados das simulações estão sintetizados na tabela 4.1, que exibe a área sob a curva ROC construída para cada medida nas diversas situações simuladas. Quanto mais próxima a área estiver de 1, maior o poder estatístico da medida.

Tabela 4.1: Área da região abaixo da curva ROC gerada para cada medida, com amostras de tamanho n

Tipo de associação	n	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	IM	CIM
Independente	10	0,50	0,50	0,48	0,45	0,50	0,50	0,34	0,34
	30	0,51	0,51	0,49	0,50	0,57	0,49	0,48	0,50
	50	0,51	0,50	0,50	0,51	0,57	0,51	0,50	0,49
Independente com outliers	10	0,87	0,89	0,56	0,52	0,48	0,57	0,03	0,33
	30	0,76	0,98	0,51	0,55	0,56	0,85	0,66	0,50
	50	0,71	1,00	0,54	0,53	0,60	0,95	0,89	0,52
Linear	10	0,80	0,76	0,75	0,72	0,69	0,61	0,40	0,52
	30	0,91	0,89	0,89	0,89	0,87	0,76	0,62	0,70
	50	0,96	0,94	0,94	0,94	0,94	0,83	0,69	0,77
Linear com outliers	10	0,86	0,95	0,72	0,75	0,78	0,86	0,06	0,64
	30	0,75	1,00	0,97	0,98	0,99	1,00	0,69	0,94
	50	0,72	1,00	1,00	1,00	1,00	1,00	0,94	0,98
Exponencial	10	0,88	0,99	0,94	0,93	0,97	0,99	0,00	0,73
	30	0,96	1,00	1,00	1,00	1,00	1,00	0,86	1,00
	50	0,99	1,00	1,00	1,00	1,00	1,00	0,90	1,00
Quadrática	10	0,16	0,71	0,16	0,14	0,90	0,97	0,54	0,52
	30	0,20	0,99	0,17	0,21	1,00	1,00	1,00	1,00
	50	0,21	1,00	0,18	0,23	1,00	1,00	1,00	1,00
Quadrática com outliers	10	0,14	0,16	0,28	0,23	0,70	0,78	0,08	0,36
	30	0,06	0,41	0,31	0,31	0,96	0,99	0,14	0,97
	50	0,05	0,64	0,31	0,32	0,99	1,00	0,94	0,99
Seno	10	0,28	0,29	0,32	0,24	0,51	0,34	0,33	0,19
	30	0,35	0,89	0,38	0,35	0,96	0,98	0,93	0,99
	50	0,40	0,98	0,42	0,42	0,99	1,00	1,00	1,00
Cúbica	10	0,34	0,45	0,32	0,28	0,43	0,54	0,37	0,30
	30	0,55	0,88	0,57	0,59	0,91	0,93	0,79	0,94
	50	0,72	0,97	0,77	0,78	0,98	0,99	0,92	0,99
Circunferência	10	0,09	0,10	0,24	0,20	0,64	0,71	0,51	0,10
	30	0,09	0,18	0,15	0,18	0,88	0,96	0,74	0,71
	50	0,09	0,38	0,15	0,18	0,95	0,99	0,96	0,94
X	10	0,09	0,03	0,14	0,11	0,00	0,66	0,42	0,02
	30	0,11	0,46	0,11	0,11	0,42	1,00	0,96	0,57
	50	0,12	0,77	0,11	0,11	0,85	1,00	1,00	0,97
Quadrado	40	0,25	0,18	0,27	0,26	0,27	0,90	0,33	0,38
	140	0,25	0,45	0,26	0,25	0,58	1,00	0,70	0,45
Correlação Local	100	0,29	1,00	0,43	0,41	0,99	1,00	1,00	1,00

4.1.1 Comparação de desempenho em cada situação simulada

Associação linear

Sob dependência linear, a medida de Pearson apresentou maior poder estatístico, seguida pela correlação de distância e pelas medidas de Spearman, Kendall, Hoeffding, HHG, CIM e IM, nessa ordem.

As diferenças de desempenho entre as medidas se tornaram perceptíveis nas simulações de amostras com ruído grande (ver descrição dos dados simulados em 3.9). Em associações lineares perfeitas ou com ruído pequeno, todas as medidas apresentaram resultados semelhantes.

Associação monotônica não linear

No caso monotônico não linear considerado nesse trabalho (ver descrição da relação exponencial em 3.9), a correlação de distância e as medidas de Spearman, Kendall, Hoeffding e HHG apresentaram maior poder estatístico, com valores correspondentes à área sob a curva ROC bastante próximos.

A correlação de Pearson, o CIM e a informação mútua apresentaram bons resultados para amostras maiores.

Associação não monotônica

Neste estudo, consideramos as seguintes relações não monotônicas funcionais: quadrática, cúbica e senóide (ver descrição das simulações em 3.9).

No caso de associação quadrática, a medida de HHG apresentou maior poder estatístico seguida pela medida D de Hoefding, correlação de distância, CIM e informação mútua. Na configuração simulada, as medidas de Pearson, Spearman e Kendall, mesmo em amostras grandes, tiveram apresentaram as respectivas curvas ROC abaixo da diagonal, isto é, curvas cujos valores da AUC são inferiores a 0,5.

No caso de associação cúbica, todas as medidas tiveram AUC abaixo de 0,5 nos testes com amostras pequenas ($n = 10$). O HHG, Hoeffding, CIM, IM e Dcor apresentaram maior poder estatístico do que as medidas de Pearson, Spearman e Kendall. Estes últimos métodos tiveram AUC superior a 0,5-0,6 apenas para amostras de tamanho 50.

Na associação senóide, novamente a AUC correspondente às medidas estudadas foi inferior a 0,5 para amostras de tamanho 10. A medida de Hoeffding apresentou maior poder estatístico, seguida pela medida de HHG, pelo CIM e pelos métodos IM e Dcor. O valor da AUC das medidas de Pearson, Spearman e Kendall esteve abaixo de 0,5, mesmo em amostras grandes.

Associação não funcional

Consideramos as seguintes associações não funcionais: circunferência, "X" e quadrado (ver descrições em 3.9) .

Na associação em forma de circunferência, a medida de HHG apresentou melhor desempenho, seguida pela medida de Hoeffding, a informação mútua e o coeficiente de informação máxima. As demais medidas não detectaram associação, isto é, apresentaram uma AUC correspondente inferior a 0,5.

Na associação em forma de "X", a medida de HHG apresentou melhor desempenho, seguida pela informação mútua e o CIM. A correlação de distância e a medida D de Hoeffding apresentaram uma curva ROC correspondente acima da diagonal apenas em amostras de tamanho 50. Novamente, as medidas de Pearson, Spearman e Kendall tiveram AUC abaixo de 0,5, mesmo em amostras grandes.

Na associação em forma de quadrado, tiveram melhor desempenho a medida de HHG e a informação mútua, sendo que a última apresentou AUC inferior a 0,5 em amostras de tamanho 40. A área sob a curva ROC das demais medidas foi próxima ou menor do que 0,5, mesmo em amostras grandes ($n = 140$).

Correlação local

No caso em que há relação linear em apenas parte das observações e não há dependência entre os demais pares observados, apresentaram bom desempenho a correlação de distância, as medidas de Hoeffding, HHG, IM e CIM.

Presença de *outliers*

Se mostraram robustas à presença de *outliers* as medidas de Spearman, Kendall, Hoeffding, HHG e CIM.

No caso independente com *outliers* (ver descrição das simulações em 3.9), se mostraram mais robustas as medidas de Spearman, Kendall e CIM, seguidas pela medida de Hoeffding. Já, a medida de HHG, apesar de se mostrar robusta nos casos de dependência entre as variáveis, teve o desempenho bastante afetado pela presença de *outliers* em amostras de tamanho 30 e 50, quando as variáveis são independentes.

4.1.2 Desempenho geral de cada medida

De acordo com as simulações realizadas, a medida de Pearson tem elevado poder estatístico no caso de dependência linear (superior ao poder das demais medidas), apresentando bons resultados mesmo para amostras pequenas ($n = 10$) e com elevado ruído (ver descrição das simulações para o caso linear). A tradicional medida não é robusta à

presença de *outliers* e teve um desempenho pior do que se escolha fosse aleatória nos casos não monotônicos e não funcionais.

De fato, pela definição da correlação de Pearson, relações descritas como a equação de uma reta devem apresentar elevado grau de correlação. Como a correlação é calculada diretamente nos valores da amostra, é esperado que a média amostral e, consequentemente, a correlação (que se baseia no produto da diferença entre os valores observados e a média) sofram grandes alterações na presença de *outliers*, conforme observamos nas simulações.

As medidas de Spearman e Kendall, de modo geral, apresentaram bom desempenho para associações monotônicas e se mostraram robustas à presença de *outliers*, o que é coerente com as definições dessas medidas, baseadas nos postos dos valores observados. Além disso, sabendo que a medida de Spearman é a medida de Pearson calculada sobre os postos, é possível compreender o bom desempenho nos casos em que há dependência linear entre os postos, ou seja, nos casos em que os valores de Y crescem quando os valores de X crescem ou nos casos em que os valores de Y descrescem quando os valores de X crescem. De forma semelhante, a correlação de Kendall, ao calcular a diferença da proporção de pares concordantes e de pares discordantes, deve apresentar bom desempenho quando ambos os valores de X e Y apenas crescem ou quando os de Y decrescem enquanto os valores de X crescem. Assim, as medidas de Speaman e Kendall apresentam comportamento semelhante e conseguem detectar dependência em associações monotônicas.

Já a correlação de distância consegue detectar, além das relações monotônicas, associações não monotônicas e não é robusta à presença de *outliers*. Conforme vimos, a correlação de distância é calculada diretamente nos valores das amostras, e, portanto, as variações nas distâncias médias decorrentes da presença de *outliers* deve provocar grandes alterações na estatística Dcor. Nos casos não funcionais simulados, a medida não apresentou bom desempenho.

A informação mútua apresentou resultados semelhantes aos da correlação de distância, mas conseguiu detectar também as relações não funcionais simuladas (o que pode ser explicado pelo cálculo da informação mútua, que utiliza a própria definição de independência estatística).

O CIM e as medidas de Hoeffding e HHG apresentaram bom desempenho nos diversos tipos de associações geradas (linear, monotônica, não monotônica e não funcional), inclusive na presença de *outliers*, de modo geral. Assim como a informação mútua, as medidas de HHG, Hoeffding e o CIM se baseiam na definição de independência, daí a generalidade de associações detectadas. O comportamento robusto na presença de *outliers* das duas primeiras medidas é esperado, como ambas se baseiam nos postos dos valores amostrais. Contudo, a medida de HHG, que, em geral, se mostrou robusta à presença de *outliers*, apresentou resultados diferentes no caso independente com ocorrência de *outliers*.

No tocante ao CIM, é possível que esta medida, por considerar diversas partições dos dados, consiga encontrar uma partição ótima em que os valores numericamente distantes dos demais não provoquem grandes perturbações.

Dentre as três medidas (CIM, HHG e Hoeffding), o CIM apresentou menor poder estatístico (é mais sensível ao ruído e ao tamanho da amostra), seguido pelas medidas de Hoeffding e de HHG.

4.1.3 Consistência dos resultados

Todas as medidas apresentaram uma AUC próxima de 0,5 no caso independente (conforme o esperado), exceto para amostras pequenas, como algumas implementações (especialmente do CIM e da informação mútua) são bastante sensíveis ao tamanho da amostra.

As simulações também se mostraram consistentes com as características do poder estatístico, estimando um poder maior para amostras maiores (exceto na presença de *outliers* para medidas não robustas a ocorrência desses elementos).

Vimos, na seção anterior 4.1.2, que os resultados podem ser em parte explicados pelas definições das medidas e se mostraram coerentes com as propriedades das mesmas. No entanto, há resultados mais desafiadores e difíceis de serem compreendidos, como o desempenho da medida de HHG no caso independente com *outliers*.

4.2 Aplicação nos dados de expressão gênica de adenocarcinomas de pulmão

Os p-valores calculados a partir dos testes de independência (realizados com cada medida estudada) entre o gene WNT5A e os 75 genes selecionados nesse estudo, após a remoção de *outliers*, constam na tabela a seguir 4.2 (tanto os originais quanto os corrigidos pelo método FDR).

Tabela 4.2: P-valores dos testes realizados com o gene WNT5A. Cada campo indica o p-valor original e o respectivo p-valor corrigido por FDR (entre parênteses). P-valores menores que 0,001 são representados por 0*.

Gene	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	IM	CIM
AES	0.329 (0.8)	0.236 (0.668)	0.237 (0.463)	0.214 (0.453)	0.061 (0.269)	0.028 (0.344)	0.397 (0.993)	0.04 (0.458)
APC	0.299 (0.8)	0.354 (0.705)	0.085 (0.316)	0.059 (0.346)	0.051 (0.257)	0.327 (0.691)	0.226 (0.856)	0.664 (0.782)
AXIN1	0.495 (0.825)	0.56 (0.799)	0.272 (0.477)	0.287 (0.495)	0.597 (0.653)	0.898 (0.968)	0.359 (0.993)	0.224 (0.61)
BCL9	0.23 (0.8)	0.274 (0.668)	0.079 (0.316)	0.112 (0.358)	0.161 (0.364)	0.23 (0.63)	0.948 (0.993)	0.405 (0.74)
BTRC	0.997 (0.997)	0.712 (0.847)	0.779 (0.823)	0.775 (0.852)	0.372 (0.58)	0.516 (0.81)	0.989 (0.995)	0.462 (0.769)
FZD5	0.864 (0.936)	0.224 (0.668)	0.024 (0.199)	0.026 (0.235)	0.039 (0.249)	0.246 (0.63)	0.864 (0.993)	0.153 (0.606)
CCND1	0.956 (0.969)	0.852 (0.94)	0.769 (0.823)	0.84 (0.852)	0.296 (0.529)	0.147 (0.55)	0.959 (0.993)	0.729 (0.792)
CCND2	0.823 (0.936)	0.905 (0.969)	0.768 (0.823)	0.821 (0.852)	0.838 (0.85)	0.981 (0.981)	0.61 (0.993)	0.943 (0.95)

CCND3	0.278 (0.8)	0.162 (0.668)	0.078 (0.316)	0.053 (0.346)	0.015 (0.164)	0.056 (0.454)	0.966 (0.993)	0.027 (0.458)
CSNK1A1	0.257 (0.8)	0.245 (0.668)	0.224 (0.463)	0.252 (0.473)	0.231 (0.433)	0.304 (0.67)	0.689 (0.993)	0.553 (0.782)
CSNK1D	0.261 (0.8)	0.264 (0.668)	0.384 (0.588)	0.305 (0.498)	0.148 (0.364)	0.089 (0.454)	0.196 (0.856)	0.259 (0.61)
CSNK1G1	0.394 (0.8)	0.209 (0.668)	0.075 (0.316)	0.062 (0.346)	0.026 (0.219)	0.02 (0.302)	0.158 (0.845)	0.181 (0.61)
CSNK2A1	0.431 (0.8)	0.266 (0.668)	0.193 (0.463)	0.194 (0.442)	0.201 (0.408)	0.512 (0.81)	0.183 (0.856)	0.608 (0.782)
CTBP1	0.599 (0.881)	0.646 (0.835)	0.955 (0.955)	0.94 (0.94)	0.738 (0.779)	0.26 (0.63)	0.577 (0.993)	0.87 (0.919)
CTBP2	0.306 (0.8)	0.327 (0.701)	0.022 (0.199)	0.02 (0.214)	0.006 (0.075)	0.089 (0.454)	0.46 (0.993)	0.146 (0.606)
CTNNB1	0.318 (0.8)	0.424 (0.726)	0.541 (0.688)	0.557 (0.72)	0.561 (0.653)	0.405 (0.723)	0.067 (0.502)	0.264 (0.61)
CTNNBIP1	0.443 (0.8)	0.426 (0.726)	0.248 (0.463)	0.297 (0.495)	0.396 (0.583)	0.713 (0.877)	0.566 (0.993)	0.105 (0.606)
CXXC4	0.412 (0.8)	0.901 (0.969)	0.427 (0.605)	0.394 (0.579)	0.493 (0.638)	0.938 (0.977)	0.15 (0.845)	0.482 (0.769)
DAAM1	0.196 (0.8)	0.403 (0.726)	0.428 (0.605)	0.441 (0.591)	0.517 (0.646)	0.911 (0.968)	0.552 (0.993)	0.545 (0.782)
DIXDC1	0.22 (0.8)	0.11 (0.668)	0.105 (0.329)	0.103 (0.358)	0.159 (0.364)	0.11 (0.454)	0.133 (0.831)	0.657 (0.782)
DKK1	0.02 (0.456)	0.009 (0.327)	0* (0.008)	0* (0.011)	0* (0.006)	0.002 (0.129)	0.012 (0.437)	0.244 (0.61)
DVL1	0.773 (0.936)	0.227 (0.668)	0.081 (0.316)	0.079 (0.346)	0.071 (0.294)	0.617 (0.833)	0.543 (0.993)	0.584 (0.782)
DVL2	0.831 (0.936)	0.18 (0.668)	0.018 (0.199)	0.012 (0.186)	0.003 (0.051)	0.032 (0.344)	0.089 (0.608)	0.088 (0.606)
EP300	0.469 (0.819)	0.154 (0.668)	0.019 (0.199)	0.028 (0.235)	0.02 (0.19)	0.112 (0.454)	0.066 (0.502)	0.048 (0.458)
FBXW11	0.602 (0.881)	0.196 (0.668)	0.012 (0.174)	0.016 (0.2)	0.047 (0.253)	0.351 (0.692)	0.228 (0.856)	0.238 (0.61)
FBXW2	0.389 (0.8)	0.481 (0.768)	0.652 (0.779)	0.658 (0.796)	0.412 (0.583)	0.257 (0.63)	0.775 (0.993)	0.615 (0.782)
FOSL1	0.266 (0.8)	0.472 (0.768)	0.241 (0.463)	0.218 (0.453)	0.133 (0.364)	0.106 (0.454)	0.559 (0.993)	0.26 (0.61)
FOXN1	0.269 (0.8)	0.151 (0.668)	0.351 (0.585)	0.296 (0.495)	0.544 (0.653)	0.212 (0.63)	0.22 (0.856)	0.723 (0.792)
FRAT1	0.308 (0.8)	0.495 (0.774)	0.369 (0.588)	0.369 (0.558)	0.462 (0.619)	0.163 (0.581)	0.49 (0.993)	0.67 (0.782)
FRZB	0.35 (0.8)	0.173 (0.668)	0.208 (0.463)	0.182 (0.44)	0.189 (0.393)	0.055 (0.454)	0.428 (0.993)	0.026 (0.458)
FSHB	0.812 (0.936)	0.698 (0.847)	0.706 (0.779)	0.707 (0.821)	0.576 (0.653)	0.617 (0.833)	0.755 (0.993)	0.373 (0.7)
FZD1	0.857 (0.936)	0.438 (0.729)	0.097 (0.316)	0.123 (0.358)	0.219 (0.42)	0.645 (0.834)	0.343 (0.993)	0.425 (0.741)
FZD2	0.024 (0.456)	0.037 (0.668)	0.12 (0.36)	0.156 (0.415)	0.4 (0.583)	0.692 (0.865)	0.511 (0.993)	0.688 (0.782)
FZD3	0.215 (0.8)	0.391 (0.726)	0.094 (0.316)	0.105 (0.358)	0.13 (0.364)	0.529 (0.81)	0.842 (0.993)	0.673 (0.782)
FZD4	0.728 (0.936)	0.723 (0.847)	0.706 (0.779)	0.827 (0.852)	0.559 (0.653)	0.555 (0.833)	0.633 (0.993)	0.036 (0.458)
FZD6	0.399 (0.8)	0.613 (0.821)	0.418 (0.605)	0.421 (0.591)	0.308 (0.537)	0.822 (0.948)	0.463 (0.993)	0.538 (0.782)
FZD7	0.199 (0.8)	0.183 (0.668)	0.096 (0.316)	0.095 (0.358)	0.334 (0.547)	0.792 (0.929)	0.05 (0.502)	0.57 (0.782)
FZD8	0.86 (0.936)	0.977 (0.979)	0.797 (0.83)	0.84 (0.852)	0.855 (0.855)	0.872 (0.968)	0.405 (0.993)	0.139 (0.606)
GSK3A	0.871 (0.936)	0.571 (0.799)	0.073 (0.316)	0.076 (0.346)	0.155 (0.364)	0.63 (0.833)	0.694 (0.993)	0.36 (0.692)
GSK3B	0.76 (0.936)	0.283 (0.668)	0.253 (0.463)	0.283 (0.495)	0.126 (0.364)	0.196 (0.63)	0.891 (0.993)	0.021 (0.458)
JUN	0.611 (0.881)	0.693 (0.847)	0.146 (0.406)	0.124 (0.358)	0.13 (0.364)	0.633 (0.833)	0.88 (0.993)	0.444 (0.757)
LEF1	0.43 (0.8)	0.812 (0.922)	0.486 (0.648)	0.432 (0.591)	0.58 (0.653)	0.951 (0.977)	0.907 (0.993)	0.097 (0.606)
LRP5	0.216 (0.8)	0.566 (0.799)	0.132 (0.38)	0.119 (0.358)	0.17 (0.364)	0.332 (0.691)	0.964 (0.993)	0.608 (0.782)
LRP6	0.244 (0.8)	0.175 (0.668)	0.501 (0.648)	0.441 (0.591)	0.422 (0.585)	0.298 (0.67)	0.186 (0.856)	0.197 (0.61)
MYC	0.073 (0.779)	0.079 (0.668)	0.67 (0.779)	0.579 (0.736)	0.406 (0.583)	0.393 (0.723)	0.026 (0.437)	0.889 (0.926)
NLK	0.506 (0.825)	0.316 (0.697)	0.05 (0.316)	0.045 (0.337)	0.056 (0.262)	0.398 (0.723)	0.744 (0.993)	0.182 (0.61)
PITX2	0.448 (0.8)	0.114 (0.668)	0* (0.014)	0* (0.016)	0* (0.01)	0.085 (0.454)	0.388 (0.993)	0.117 (0.606)
PORCN	0.011 (0.403)	0.051 (0.668)	0.234 (0.463)	0.195 (0.442)	0.147 (0.364)	0.115 (0.454)	0.034 (0.437)	0.066 (0.55)
PPP2CA	0.831 (0.936)	0.627 (0.824)	0.274 (0.477)	0.321 (0.513)	0.601 (0.653)	0.916 (0.968)	0.905 (0.993)	0.268 (0.61)
PPP2R1A	0.829 (0.936)	0.25 (0.668)	0.056 (0.316)	0.071 (0.346)	0.032 (0.24)	0.085 (0.454)	0.484 (0.993)	0.236 (0.61)
PYGO1	0.901 (0.939)	0.975 (0.979)	0.652 (0.779)	0.653 (0.796)	0.599 (0.653)	0.834 (0.948)	0.283 (0.95)	0.62 (0.782)
SENP2	0.57 (0.872)	0.953 (0.979)	0.679 (0.779)	0.739 (0.828)	0.641 (0.687)	0.763 (0.923)	0.482 (0.993)	0.3 (0.643)
SFRP1	0.938 (0.964)	0.843 (0.94)	0.702 (0.779)	0.74 (0.828)	0.567 (0.653)	0.526 (0.81)	0.793 (0.993)	0.561 (0.782)
SFRP4	0.891 (0.939)	0.064 (0.668)	0.002 (0.052)	0.001 (0.03)	0* (0.006)	0.005 (0.19)	0.035 (0.437)	0.009 (0.458)
SHFM3	0.44 (0.8)	0.748 (0.863)	0.6 (0.75)	0.676 (0.805)	0.475 (0.625)	0.79 (0.929)	0.77 (0.993)	0.175 (0.61)
SLC9A3R1	0.811 (0.936)	0.592 (0.808)	0.422 (0.605)	0.598 (0.747)	0.455 (0.619)	0.171 (0.583)	0.056 (0.502)	0.416 (0.741)
SOX17	0.748 (0.936)	0.978 (0.979)	0.3 (0.512)	0.333 (0.521)	0.211 (0.417)	0.571 (0.833)	0.745 (0.993)	0.353 (0.692)
T	0.071 (0.779)	0.042 (0.668)	0.085 (0.316)	0.138 (0.384)	0.101 (0.364)	0.089 (0.454)	0.935 (0.993)	0.332 (0.673)
TCF7	0.874 (0.936)	0.72 (0.847)	0.871 (0.882)	0.807 (0.852)	0.768 (0.789)	0.584 (0.833)	0.714 (0.993)	0.477 (0.769)
TCF7L1	0.25 (0.8)	0.173 (0.668)	0.187 (0.463)	0.178 (0.44)	0.163 (0.364)	0.104 (0.454)	0.693 (0.993)	0.634 (0.782)
TLE1	0* (0.016)	0.004 (0.324)	0.003 (0.052)	0.004 (0.067)	0.003 (0.051)	0.008 (0.19)	0.034 (0.437)	0.146 (0.606)
TLE2	0.63 (0.891)	0.16 (0.668)	0.161 (0.432)	0.11 (0.358)	0.04 (0.249)	0.014 (0.27)	0.291 (0.95)	0.148 (0.606)
WIF1	0.352 (0.8)	0.569 (0.799)	0.378 (0.588)	0.245 (0.47)	0.076 (0.301)	0.231 (0.63)	0.995 (0.995)	0.135 (0.606)
WISP1	0.715 (0.936)	0.979 (0.979)	0.869 (0.882)	0.822 (0.852)	0.761 (0.789)	0.975 (0.981)	0.587 (0.993)	0.915 (0.94)
WNT1	0.07 (0.779)	0.076 (0.668)	0.079 (0.316)	0.096 (0.358)	0.167 (0.364)	0.368 (0.707)	0.426 (0.993)	0.197 (0.61)

WNT11	0.376 (0.8)	0.562 (0.799)	0.471 (0.642)	0.405 (0.584)	0.39 (0.583)	0.498 (0.81)	0.818 (0.993)	0.822 (0.881)
WNT16	0.296 (0.8)	0.357 (0.705)	0.205 (0.463)	0.161 (0.415)	0.144 (0.364)	0.342 (0.692)	0.631 (0.993)	0.95 (0.95)
WNT2	0.488 (0.825)	0.388 (0.726)	0.36 (0.587)	0.262 (0.479)	0.141 (0.364)	0.239 (0.63)	0.767 (0.993)	0.684 (0.782)
WNT2B	0.209 (0.8)	0.235 (0.668)	0.176 (0.454)	0.208 (0.453)	0.125 (0.364)	0.492 (0.81)	0.894 (0.993)	0.58 (0.782)
WNT3	0.724 (0.936)	0.353 (0.705)	0.095 (0.316)	0.069 (0.346)	0.047 (0.253)	0.289 (0.67)	0.783 (0.993)	0.049 (0.458)
WNT4	0.551 (0.861)	0.677 (0.847)	0.232 (0.463)	0.231 (0.468)	0.347 (0.554)	0.69 (0.865)	0.835 (0.993)	0.589 (0.782)
WNT5B	0.289 (0.8)	0.426 (0.726)	0.686 (0.779)	0.711 (0.821)	0.336 (0.547)	0.594 (0.833)	0.798 (0.993)	0.289 (0.637)
WNT6	0.52 (0.83)	0.288 (0.668)	0.494 (0.648)	0.372 (0.558)	0.296 (0.529)	0.255 (0.63)	0.019 (0.437)	0.246 (0.61)
WNT7A	0.151 (0.8)	0.294 (0.668)	0.451 (0.626)	0.463 (0.609)	0.517 (0.646)	0.455 (0.794)	0.542 (0.993)	0.325 (0.673)
WNT7B	0.328 (0.8)	0.575 (0.799)	0.245 (0.463)	0.243 (0.47)	0.316 (0.538)	0.879 (0.968)	0.279 (0.95)	0.726 (0.792)

As associações detectadas, com $\alpha = 0,05$, são exibidas nas tabelas 4.3 e 4.4, considerando os p-valores não ajustados e ajustados, respectivamente.

Tabela 4.3: Associações detectadas com p-valores não corrigidos

Gene	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	IM	CIM
AES						X		X
FZD5			X	X	X			
CCND3					X			X
CSNK1G1					X	X		
CTBP2			X	X	X			
DKK1	X	X	X	X	X	X	X	
DVL2			X	X	X	X		
EP300			X	X	X			X
FBXW11			X	X	X			
FRZB								X
FZD2	X	X						
FZD4								X
GSK3B								X
MYC							X	
NLK			X	X				
PITX2			X	X	X			
PORCN	X						X	
PPP2R1A					X			
SFRP4			X	X	X	X	X	X
T		X						
TLE1	X	X	X	X	X	X	X	
TLE2					X	X		
WNT3					X			X
WNT6							X	

Tabela 4.4: Associações detectadas após correção dos p-valores por FDR

Gene	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	IM	CIM
DKK1			X	X	X			
PITX2			X	X	X			
SFRP4				X	X			
TLE1	X							

Após correção por FDR, foram obtidos p-valores menores que 0,05 em testes com os genes DKK1, PITX2, SFRP4 e TLE1.

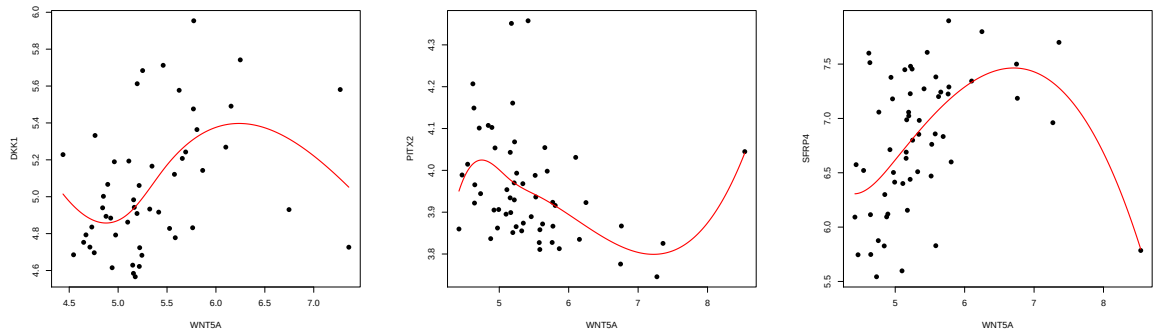
A medida de Pearson detectou associação apenas com o gene TLE1 que, como observamos em 4.1, é linear, conforme o esperado, de acordo com os resultados das simulações. Já a medida de Spearmam detectou duas associações (DKK1 e PITX2) e a medida de Kendall detectou três associações (DKK1, PITX2 e SFRP4), sendo que nenhuma das duas medidas identificou associação com o TLE1.

Mesmo após a remoção dos *outiers*, pode-se perguntar se o teste de independência entre os genes WNT5A e TLE1 com a medida de Pearson foi afetado pela presença de alguns valores numericamente distantes ainda presentes na amostra. Para responder essa pergunta, é preciso considerar que o poder estatístico da medida de Pearson no caso linear é superior ao das demais medidas (conforme verificamos nas simulações). Além disso, os p-valores originalmente obtidos com as medidas de Spearman e Kendall foram pequenos (0,003 e 0,004, repectivamente), o que nos leva a acreditar que a medida de Pearson detectou corretamente a associação.

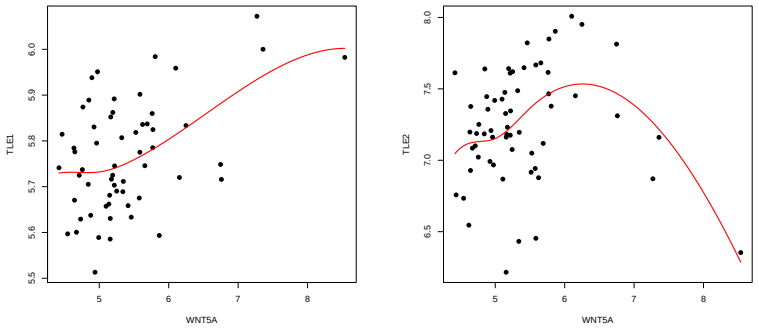
No tocante às associações detectadas pela medida de Kendall ou de Spearman, e não identificadas pela medida de Pearson, é possível que as associações com os genes PITX2 e SFRP4 não sejam lineares, uma vez que os testes com a medida de Pearson apresentraram p-valores não corrigidos grandes (0,448 e 0,891, respectivamente). Não podemos dizer o mesmo da associação com o gene DKK1, pois, apesar do p-valor corrigido do teste com a medida de Pearson ser maior que 0,05, o p-valor original foi pequeno (0,02).

A medida D de Hoeffding detectou, com os p-valores corrigidos, as mesmas associações identificadas pela medida de Kendall. Já, com os p-valores originais, foram encontradas cinco associações além de outras nove encontradas também pelas medidas de Spearman e Kendall.

A figura a seguir 4.1 ilustra os gráficos de dispersão entre o WNT5A e alguns genes com os quais foram detectadas associações, as quatro primeiras com os p-valores corrigidos, e a última com p-valor não corrigido.



(a) DKK1 (b) PITX2 (c) SFRP4



(d) TLE1 (e) TLE2

Figura 4.1: Gráficos de dispersão dos níveis de expressão gênica (em escala logarítmica) entre o WNT5A e cinco genes associados, de amostras de adenocarcinoma de pulmão (estágio I), após a remoção de *outliers*. Exibimos apenas 63 das 104 amostras consideradas nos testes, que correspondem às amostras do *Memorial Sloan-Kettering Cancer Center*.

Quando comparamos os p-valores não corrigidos de todas as medidas estudadas, podemos verificar resultados, em grande parte, semelhantes aos obtidos nas simulações. Notamos, por exemplo, a semelhança de desempenho entre as medidas de Spearman e Kendall; o maior número de associações detectadas pela medida de Hoeffding quando comparada a outras tradicionais; e a proporção menor de associações detectadas pela correlação de Pearson, indicando possíveis associações não lineares.

Nossas simulações indicam que associações não detectadas pelas medidas de Spearman e Kendall são não monotônicas ou correlações locais. De fato, observando o gráfico de dispersão dos níveis de expressão gênica do WNT5A e do TLE2, com o qual foi identificada associação pelas medidas de Hoeffding e HHG, mas que, nos testes de independência com as medidas de Spearman e Kendall, apresentou p-valor grande, a associação parece ser não monotônica.

Contudo, diferentemente do que observamos nas simulações, o HHG não detectou mais associações do que as demais medidas. Além disso, todos os métodos implementados com

bootstrap apresentaram p-valores maiores e não detectaram associação após a correção por FDR.

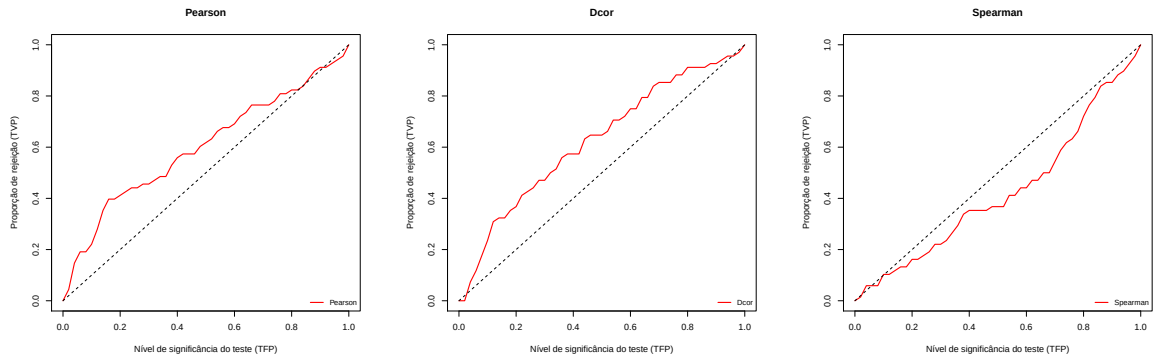
É possível que os resultados desses testes não tenham convergido, uma vez que verificamos que houve diferenças significativas entre os valores obtidos com 1000 permutações ($B = 1000$) e os valores obtidos com 10000 permutações ($B = 10000$). Erros da ordem de 10^{-2} podem ter contribuido para que os p-valores ajustados ficassem grandes. Testes com mais de 10000 reamostragens seriam inviáveis dentro do prazo estabelecido para o desenvolvimento do trabalho, de acordo com o que observamos pelo tempo de execução do programa que calcula o CIM.

Quando consideramos os p-valores não-corrigidos, boa parte das associações são detectadas tanto por métodos implementados com *bootstrap*, quanto por métodos que utilizam uma fórmula analítica de distribuição de probabilidade.

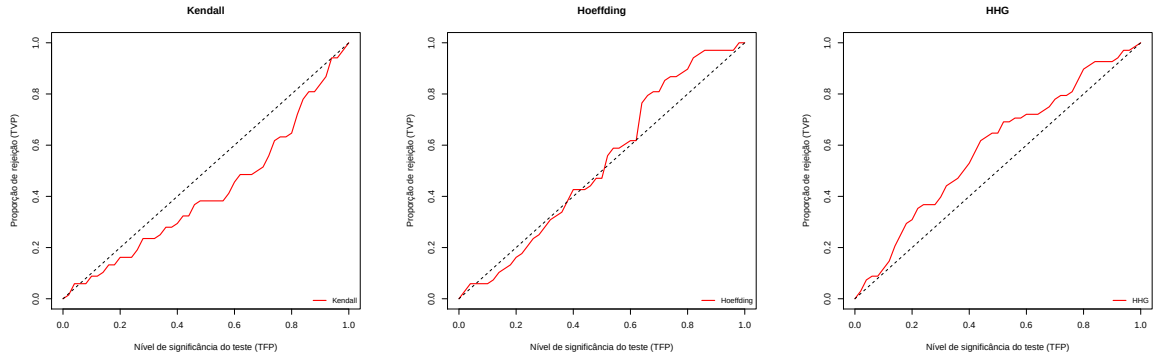
4.2.1 Validação do experimento

Para validar o experimento, realizamos testes de independência com 68 genes de controle, advindos de outros organismos, que não devem estar relacionados aos demais genes. Na figura 4.2, são exibidas as curvas ROC contruídas a partir dos testes com os 68 genes de contole.

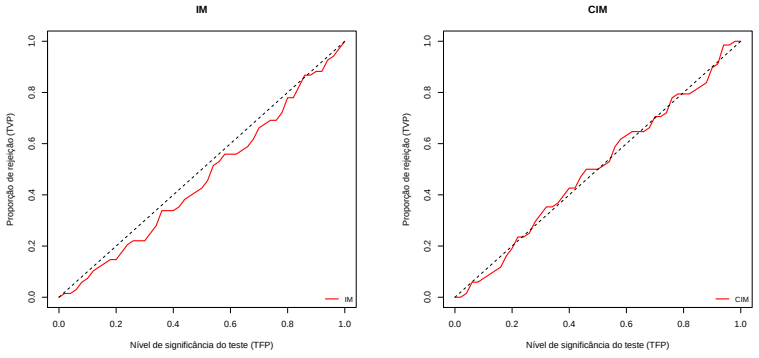
Neste cenário, a curva ROC ideal estaria na diagonal do quadrado unitário. Observamos que as medidas de Pearson, Dcor e HHG apresentaram curvas um pouco mais distantes da diagonal do que as demais, o que é aceitável, pois os testes com as duas primeiras medidas são significativamente afetados pela presença de *outliers* na amostra e os testes com a medida de HHG também o são, no caso independente, segundo as simulações realizadas.



(a) Pearson (b) Correlação de distância (c) Spearman



(d) Kendall (e) Hoeffding (f) HHG



(g) IM (h) CIM

Figura 4.2: Curvas ROC construídas a partir dos resultados dos testes de independência entre o gene WNT5A e os genes de controle

Capítulo 5
Conclusões

Em nosso estudo comparativo das medidas de Pearson, Spearman, Kendall, Hoeffding, HHG, IM, CIM e Dcor, verificamos diferenças significativas de suas potencialidades e limitações. Tais resultados, de modo geral, estiveram consistentes com nossas expectativas, de acordo com o conhecimento prévio de cada método.

A tabela a seguir 5.1 sintetiza as propriedades verificadas.

Tabela 5.1: Tipo de associações detectadas por cada medida.

Tipos de Associação	Pearson	Dcor	Spearman	Kendall	Hoeffding	HHG	IM	CIM
Linear	X	X	X	X	X	X	X	X
Monotônica não linear		X	X	X	X	X	X	X
Não monotônica		X			X	X	X	X
Não funcional					X	X	X	X
Robusta à presença de outliers			X	X	X	X (com exceções)		X

Apesar das particularidades de cada situação simulada, nosso estudo revelou que, de modo geral, a medida de HHG apresentou maior poder estatístico, seguida pela medida D de Hoeffding. No caso de dependência linear, a tradicional medida de Pearson superou as demais.

Após a caracterização de cada medida, baseada nas simulações, ilustramos o uso dos métodos estudados em dados de expressão gênica de câncer de pulmão. Nesse contexto, verificamos alguns comportamentos já identificados pelo nosso estudo. Contudo, os métodos implementados com *bootstrap* apresentaram, nos testes realizados com os dados biológicos, resultados não reproduzidos nas simulações. Apesar do estudo comparativo apontar a medida de HHG como a de maior poder estatístico, quando corrigidos os p-valores, a mesma não detectou nenhuma associação nos testes com os dados biológicos, enquanto outras como as medidas de Hoeffding, Spearman e Kendall identificaram dependências. Além disso, esta mesma medida, apresentou alterações de desempenho no caso de independência entre as variáveis e ocorrência de *outliers* nas amostras, apesar de sua definição sugerir um comportamento robusto à presença de *outliers*.

Não obstante essas limitações, foi possível estabelecer uma caracterização geral e consistente das diversas medidas de dependência consideradas nesse estudo, proporcionando

ao pesquisador informações relevantes para a escolha do método estatístico a ser utilizado em seu trabalho.

Referências Bibliográficas

[1] Director's Challenge Consortium for the Molecular Classification of Lung Adenocarcinoma *et al*. Gene expression–based survival prediction in lung adenocarcinoma: a multi-site, blinded validation study. *Nature Medicine*, **14**:822–827, 2008.

[2] Fujita A. *Análise de dados de expressão gênica: normalização de microarrays e modelagem de redes regulatórias*. Tese de Doutoramento, Universidade de São Paulo, 2007.

[3] Fujita A, Sato JR, Demasi MA, Sogayar MC, Ferreira CE e Miyano. Comparing Pearson, Spearman and Hoeffding's D measure for gene expression association analysis. *Journal of Bioinformatics and Computational Biology*, **7**(4):663–684, 2009.

[4] Soper HE, Young AW, Cave BM, Lee A e Pearson K. On the distribution of the correlation coefficient in small samples. Appendix II to the papers of "Student" and R. A. Fisher. A cooperative study. *Biometrika*, **11**(4):328–413, 1917.

[5] Spearman C. The proof and measurement of association between two things. *The American Journal of Psychology*, **15**(1):72–101, 1904.

[6] Hoeffding W. A non-parametric test of independence. *The Annals of Mathematical Statistics*, **19**:546–557, 1948.

[7] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Mitzenmacher M Lander ES e Sabeti PC. Detecting Novel Associations in Large Data Sets. *Science*, **334**(6062):1518–1524, 2011.

[8] Steuer R, Kurths J, Daub CO, Weise J e Selbig J. The mutual information: detecting and evaluating dependencies between variables. *Bioinformatics*, **18**:S231–S240, 2002.

[9] Simon N e Tibshirani R. *Comment on "Detecting Novel Associations in Large Data Sets" by Reshef et al., Science Dec. 16, 2011*, 2011. `http://www-stat.stanford.edu/~tibs/reshef/comment.pdf`.

[10] Szekely G, Rizzo M e Bakirov N. Measuring and testing independence by correlation of distances. *The Annals of Statistics*, **35**:2769–2794, 2007.

[11] Gorfine M, Heller R e Heller Y. *Comment on "Detecting Novel Associations in Large Data Sets"*. `http://ie.technion.ac.il/~gorfinm/files/science6.pdf`.

[12] Heller R, Heller Y e Gorfine M. A consistent multivariate test association based on ranks of distances. *Biometrika*. (aceito). http://arxiv.org/abs/1201.3522.

[13] Kendall M. A new measure of rank correlation. *Biometrika*, **30**(1-2):81–93, 1938.

[14] Efron B. Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, **7**(1):1–26, 1979.

[15] Benjamini Y e Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, **57**(1):289–300, 1995.

[16] Fawcett T. An introduction to ROC analysis. *Pattern Recognition Letters*, **27**:861–874, 2006.

[17] Zaha A, Ferreira HB e Passaglia LMP *et al*. *Biologia molecular básica*. Mercado Aberto, 3ª edição, 2003.

[18] R Development Core Team. *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria, 2010. ISBN 3-900051-07-0. http://www.R-project.org/.

[19] Rizzo ML e Szekely GJ. *energy: E-statistics (energy statistics)*, 2011. R package version 1.4-0. http://CRAN.R-project.org/package=energy.

[20] Harrell FE Jr *et al.*. *Hmisc: Harrell Miscellaneous*, 2012. R package version 3.9-3. http://CRAN.R-project.org/package=Hmisc.

[21] Hausser J e Strimmer K. *entropy: Entropy and Mutual Information Estimation*, 2012. R package version 1.1.7. http://CRAN.R-project.org/package=entropy.

[22] Reshef D e Reshef Y. *MINE: Maximal Information-based Nonparametric Exploration*. http://www.exploredata.net.

[23] *cArray - Array data management system*. https://array.nci.nih.gov/caarray/project/details.action?project.experiment.publicIdentifier=jacob-00182.

[24] Mazieres J, He B, You L, Xu Z e Jablons DM. Wnt signaling in lung cancer. *Cancer letters*, **222**:1–10, 2005.

[25] Gautier L, Cope L, Bolstad BM e Irizarry RA. affy—analysis of Affymetrix GeneChip data at the probe level. *Bioinformatics*, **20**(3):307–315, 2004.

[26] Laboratory of Population Genetics, National Cancer Institute, National Institutes of Health. *Custom Chip Definition Files (CDF) for Unified Gene Expression Analysis*. `http://masker.nci.nih.gov/ev/`.

[27] American Joint Committee on Cancer. *AJCC Cancer Staging Manual*, capítulo 19, páginas 167–177. Springer, 6ª edição, 2002. `http://www.cancerstaging.org/products/csmanual6ed-3.pdf`.

[28] Bussab WO e Morettin PA. *Estatística Básica*. Editora Saraiva, 5ª edição, 2002.

[29] Fisher RA. *Statistical Methods for Research Workers*. Oliver and Boyd, Edinburgh, 1932.

[30] Wikipedia. *Gene expression — Wikipedia, The Free Encyclopedia*, 2012. `http://en.wikipedia.org/wiki/Gene_expression`.

[31] Wikipedia. *Microarranjo — Wikipédia, a enciclopédia livre*, 2012. `http://pt.wikipedia.org/wiki/Microarranjo`.

[32] Ferreira Filho D e Leandro RA. *Análise de Microarray usando o R e o Bioconductor*, 2009. Tutorial apresentado no 54º RBRAS e 13º SEAGRO. `http://www.lce.esalq.usp.br/roseli/TUTORIAL_MICROARRAY.pdf`.

Parte II

Parte Subjetiva

Realizei este projeto como uma iniciação científica, sob orientação do professor André Fujita, com apoio financeiro do PIBIC/CNPq, apresentando, em outubro, este trabalho no 20º SIICUSP.

Esta iniciação, que também é meu trabalho de conclusão do curso de Bacharelado em Ciência da Computação, representa minha primeira experiência com um projeto de pesquisa e com a análise, ainda que simplesmente ilustrativa, de dados "reais". Neste projeto, pude concretizar a vontade de, com base no que aprendi na graduação, fazer algo de "verdade", isto é, trabalhar num problema com aplicações no mundo real. Reuni num só trabalho fundamentos estatísticos, ferramentas computacionais e aplicações em dados biológicos.

As relações do meu trabalho com o curso do BCC (Bacharelado em Ciência da Computação), bem como as considerações sobre os principais desafios encontrados e o futuro do trabalho, serão abordadas nas próximas seções.

Desafios e frustrações

No início de 2012, quando tive minha primeira reunião com supervisor deste trabalho, o professor André, lembrava-me pouco de Estatística e estava ainda no "clima" de férias, período em que costumamos deixar muitos conhecimentos adormecidos.

A primeira dificuldade foi sair desse "clima" e começar a desenvolver algo neste trabalho. Sabia que precisava começar logo, afinal, no período letivo, o tempo parece extremamente curto.

Dados os passos iniciais, o restante parecia fluir bem, com a a constante interação com o orientador.

A proposta do projeto foi aumentando à medida que outros testes de independência estatística foram adicionados ao estudo. Enquanto isso, simulações e mais simulações estavam sendo realizadas.

Uma das principais dificuldades nesse trabalho foi realizar simulações muito demoradas, como os testes de independência feitos com o Coeficiente de Informação Máxima (CIM). Esses testes precisaram ser realizados numerosas vezes, algumas por erros decorrentes de execução paralela das simulações, alterações no número de permutações realizadas nos testes estatísticos, ou mesmo por mudanças na implementação do teste de independência. As simulações demoravam dias, e, numerosas vezes, por algum motivo, foram interrompidas. Então, tive que estar sempre "supervisionando" as simulações para perceber essas não raras interrupções.

Com a entrada dos dados "reais", as dificuldades foram aumentando. Tive que aprender a processar os dados e como obter os valores correspondentes aos níveis de expressão

gênica. No começo foi um pouco difícil aprender usar o pacote *affy* do *Bioconductor* para processar os dados biológicos, pois achava a documentação insuficiente e havia muitos termos novos para mim, mesmo palavras chaves nesse contexto, como *probes* e *probesets*.

Depois que finalmente consegui processar os dados e comecei a entender como obter a expressão gênica, me deparei com um novo desafio. Precisava de um valor para cada gene, mas os *probes* (ou "sondas") não estavam agrupados por gene. Poderia haver mais de um grupo (*probeset*) associado a um único gene. Encontrei artigos tratando dessa questão, entre eles, alguns afirmando que se deve escolher um *probeset* para representar um gene e outros falando dos problemas envolvidos na escolha de um *probeset*. Por fim, percebi que eu poderia utilizar um pacote do R para fazer um outro tipo de "agrupamento" das *probes*, que era por gene. Foram longas horas na *internet* para encontrar o que eu precisava e aprender como usar.

Outro desafio nesse trabalho foi a análise dos dados biológicos e minha inexperiência com estatística. Várias questões foram surgindo, como o fato das amostras serem de laboratórios diferentes e a necessidade de remover o chamado "efeito de sítio".

Durante todo processo, tive que lidar com vários artigos de Estatística, Bioinformática e até de Medicina. Muitas vezes eram apenas consultas rápidas, mas que acabavam se tornando demoradas, como alguns desses artigos utilizam uma linguagem muito técnica e específica da área. Além disso, nesse período todo, grandes foram os esforços para me concentrar nas questões acadêmicas em detrimento de outras preocupações, e para vencer o sono acumulado. Fiquei preocupada quando percebi que estava comemorando quando tinha três ou quatro horas de sono, em alguns períodos do ano. Nesses momentos, alguns feriados e semanas do *break* ajudaram bastante.

Minha maior frustração nesse trabalho foi ter deixado a análise mais detalhada dos resultados tão próxima da entrega e não ter conseguido ilustrar tudo o que eu gostaria na análise dos dados biológicos. Acho que eu poderia ter feito uma análise melhor e mais cautelosa, se ela fosse feita num prazo maior.

Paralelo entre o trabalho de formatura e as disciplinas do BCC

Não foram raras as vezes que me perguntaram, enquanto explicava sobre meu trabalho de conclusão de curso e iniciação científica, o que tinha de computação nisso tudo.

Nessas ocasiões, tento explicar que esse trabalho é uma forma de reunir, além dos conhecimentos "diretos" em estatística e matemática adquiridos no BCC, várias outras capacidades desenvolvidas ao longo do curso, algumas mais gerais, como aprender a aprender, aprender a ler textos mais difíceis, enfim, aprender a se virar, e outras mais

específicas, como processar os dados no computador e fazer simulações, utilizando habilidades de "computeiros".

Embora, em princípio, fazer simulações e utilizar pacotes do R para processar dados biológicos possa em parte ser realizado por pessoas que não são da área de computação, o trabalho seria muito mais árduo, não fossem os conhecimentos computacionais, que me permitiram automatizar e verificar o que fiz no trabalho. Além disso, executei simulações em paralelo e pude identificar o problema de concorrência, graças ao que aprendi no IME.

Um "bcçóide" tem muito pelo que se interessar nesse trabalho, pois ele percebe o papel fundamental do computador nas análises, e, com um pouco de interesse em estatística e biologia, motivar-se-á pelo propósito e aplicações do mesmo.

Todas as disciplinas cursadas no BCC trouxeram algum amadurecimento que direta ou indiretamente contribuiram para o desenvolvimento desta iniciação científica.

A seguir listarei aquelas que julgo mais importantes para este trabalho.

- **MAE0121 - Introdução a Probabilidade e a Estatística I**
  MAE0212 - Introdução à Probabilidade e à Estatística II
  Disciplinas fundamentais para compreender o objeto de estudo do trabalho e também as técnicas estatísticas utilizadas.

- **MAC0460 - Aprendizagem Computacional: Modelos, Algoritmos e Aplicações**
  Por abordar várias técnicas estatísticas e computacionais, essa disciplina tem bastante relação com o meu trabalho e com a área da Bioinformática na qual ele se insere. Técnicas como *bootstrap* e *curvas ROC*, que foram utilizadas nesse trabalho, foram ensinadas nessa disciplina.

- **MAC0110 - Introdução à Computação**
  MAC0122 - Princípios de Desenvolvimento de Algoritmos
  MAC0323 - Estruturas de Dados
  As três disciplinas foram fundamentais para que eu aprendesse a programar e conseguisse escrever os códigos dos testes de independência, das simulações, da geração das curvas ROC e pudesse automatizar o processo de análise dos dados biológicos.

- **MAC0338 Análise de Algoritmos**
  Nessa disciplina, aprendi conceitos de complexidade computacional importantes para o trabalho, uma vez que as simulações eram demoradas e precisei me preocupar em fazer códigos mais eficientes.

- **MAC0211 - Laboratório de Programação I**
  MAC0242 - Laboratório de Programação II
  Nessas disciplinas, tive contato com linguagens *script*, como *perl* e *python*, o que

facilitou o desenvolvimento no ambiente R. Além disso, em Laboratório de Programação I, aprendi um pouco mais de *bash script*, o que me auxiliou em várias operações do trabalho.

- **MAC0316 - Conceitos Fundamentais de Linguagens de Programação**
  MAC0319 - Programação Funcional Contemporânea
  Nessas disciplinas, aprendi a fazer programas mais curtos e adquiri uma certa experiência que me ajudaram a programar mais eficientemente no ambiente R.

- **MAT0138 Álgebra I para Computação**
  Essa disciplina me forneceu uma importante base matemática para a compreensão dos artigos sobre as medidas de dependência estudadas.

- **MAT0111 - Cálculo Diferencial e Integral I**
  Fundamental para a compreensão dos artigos sobre medidas de dependência e dos fundamentos estatísticos. Utilizei conhecimentos de cálculo I diretamente no trabalho para calcular a área sob a curva ROC das simulações.

- **MAT0139 - Álgebra Linear para Computação**
  Importante para a compreensão dos artigos sobre medidas de dependência, especialmente as que utilizam o conceito de distância.

- **MAC0438 - Programação Concorrente**
  Importante para entender os problemas em execuções paralelas que compartilham recursos, como aconteceu nesse trabalho.

Trabalhos futuros

Uma forma de enriquecer o projeto é criar mais situações a serem simuladas a fim de explorar ainda mais os diferentes comportamentos das medidas de dependência estudadas.

Poderíamos, também, realizar testes com outras implementações da informação mútua, e buscar outras medidas de dependência a serem incluídas no estudo.

Para enriquecer o exemplo com os dados biológicos poderíamos continuar a buscar mais associações entre os genes que pudéssemos classificar como não monotônicas.

Pretendemos dar continuidade ao trabalho submetendo um artigo com os resultados obtidos. Além disso, pretendo continuar na área de Bioinformática e Estatística Computacional, prosseguindo meus estudos na pós-graduação.

Considerações finais

Superação de limites e grandes desafios são palavras que marcam a vida de um "bcçóide" no IME. Os anos de BCC ensinam a ter força de vontade e a enfrentar diversas situações.

Para mim, em especial, o BCC representa uma enorme superação. Desde o vestibular, superando as dificuldades em ter concentração e administrar o tempo de estudo, às disciplinas, varando noites para concluir trabalhos. Ao entrar na terceira chamada do vestibular, tive medo de não acompanhar o curso, que sabia ser puxado e ter disciplinas desafiadoras.

De certo modo, esse medo foi bom, pois me ajudou a estudar com mais cautela. Contudo, fui percebendo que não era preciso ser um "gênio" para fazer BCC. Era preciso, sobretudo, ter esforço e dedicação.

Muitas vezes fiquei perdida nas conversas "bcçóides", por não ter a mesma cultura *nerd* de tantos outros colegas. Mas percebi que existem "bcçóides" e "bcçóides" no IME e uma grande diversidade de interesses. Nessa diversidade que encontrei minha identificação com o curso.

Se eu voltasse no tempo e prestasse vestibular novamente, faria o mesmo curso e o mesmo TCC. Mas, mudaria muitas coisas: me esforçaria para acompanhar melhor as aulas, dada a minha inibição em participar das mesmas e minha tendência a me distrair facilmente; e faria de tudo para aproveitar mais a companhia dos meus colegas.

Apesar de algumas lacunas que eu acredito ter deixado na graduação, tenho a convicção de que recebi, no BCC no IME, uma formação sólida para enfrentar e aprender a enfrentar variados problemas e de que a iniciação científica realizou um papel chave no meu amadurecimento para esses desafios.

Um muito obrigada a todos os funcionários, professores e colegas, sem os quais fazer este curso não seria possível, e ao professor André pela oportunidade de realizar este trabalho e por toda a ajuda.

É difícil dizer que algo acabou. A finalização deste trabalho e, possivelmente, deste curso é um novo começo...