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.

São Paulo, 2012

Sumário

Resumo

Palavras-chave:

Parte I

Parte Objetiva

Capítulo 1
Introdução

Nesta seção farei uma contextualização da área e escreverei sobre a motivação para o trabalho.

1.1 Objetivos

1.1.1 Objetivos gerais

Os objetivos do presente trabalho são um estudo comparativo entre diversas medidas de dependência entre variáveis aleatórias. São elas as medidas de correlação de Pearson [11] e Spearman [12], o tau de Kendall [9], a informação mútua [10], a medida D de Hoeffiding [7], o Coeficiente de Informação Máxima (CIM) [2], a correlação de distância [3] e a medida de Heller, Heller e Gorfine (HHG) [4], com posterior aplicação em dados biológicos reais.

1.1.2 Objetivos específicos

1. Estudar comparativamente em dados simulados as medidas de Pearson e Spearman, o tau de Kendall, a informação mútua, a medida D de Hoeffding, o CIM, a correlação de distância e a medida de HHG de forma a identificar sob quais condições cada medida é capaz de detectar dependência.

2. Realizar aplicações em dados biológicos de expressão gênica advindos de tecnologia de microarranjos de *DNA*.

1.2 Estrutura do presente trabalho

Nesta seção, descreverei os tópicos a serem abordados.

Capítulo 2
Fundamentação teórica

Nesta seção, farei uma breve abordagem sobre os tópicos abaixo e outros que julgar pertinentes no desenvolvimento da monografia.

2.1 Dependência entre variáveis aleatórias

2.2 Teste de hipóteses

2.2.1 *P*-valor

2.2.2 *Bootstrap*

2.3 Curva *ROC*

2.4 Expressão gênica e microarranjos de *DNA*

2.5 O método *Fast Cyclic Loess* de normalização

Capítulo 3
Ferramentas utilizadas

Nessa seção listarei as principais ferramentas utilizadas para o desenvolvimento do trabalho.

Capítulo 4
Metodologia

O estudo comparativo se baseia na avaliação do poder estatístico das medidas de dependência em diversos tipos de dados gerados com a ferramenta R [6]. 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

Realizamos o teste de independência com cada uma das medidas estudadas nesse trabalho e posteriormente comparamos o poder estatístico das mesmas.

Dada a significância do teste, o poder estatístico é estimado pela proporção de vezes que a hipótese nula é rejeitada em 1000 simulações.

Variando a significância do teste de 0 a 1, podemos construir uma curva do poder estimado e interpretá-la como uma curva ROC, tomando o poder como a taxa de verdadeiros positivos e a significância do teste 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 produzida, que nos fornece uma medida geral do desempenho de cada método na identificação de associação entre variáveis. 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.

4.1 Correlação de Pearson

O coeficiente de correlação de Pearson é uma medida de correlação entre duas variáveis aleatórias. É bastante 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 **y** diminuem à medida que os valores de **x** aumentam. Se r_p vale 0, não há correlação linear. Quando r_p vale 1, os valores de **y** aumentam quando os valores de **x** aumentam e há uma reta que passa por todos os pontos (x_i, y_i). Valores intermediários podem, em certas situações, indicar o quão forte é a associação entre **x** e **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.

4.2 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 [1].

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 **x** e **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 **y** crescem quando **x** cresce, ou se os valores de **y** diminuem, à medida que os valores de **x** aumentam.

Como são utilizados os postos no lugar dos dados, a presença de *outliers* [2] não prejudica

[1] O posto de um elemento de um vetor é a posição do mesmo no vetor com os dados em ordem crescente.
[2] Um *outlier* é uma observação numericamente distante das demais na amostra.

a identificação de dependência entre variáveis.

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.

4.3 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.

4.4 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:

$$I(X,Y) = \int_Y \int_X f(x,y) \log \frac{f(x,y)}{f(x)f(y)} dxdy,$$

onde $f(x,y)$ é a função de densidade de probabilidade conjunta de X e Y e $f(x)$ e $f(y)$ 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 $p(x,y)$ é a função de distribuição de probabilidade conjunta de X e Y e $p(x)$ e $p(y)$ 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. Como não se conhece uma fórmula fechada da distribuição de probabilidade dessa medida, construímos o teste de independência com base na técnica computacional de *bootstrap*.

4.5 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 $F(x, y)$ a função de distribuição acumulada conjunta de X e Y e $G(x)$ e $H(y)$ as funções de distribuição acumulada marginais de X e Y, respectivamente. D é uma medida da distância entre $F(x, y)$ e $G(x)H(y)$.

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

$$F(x, y) = G(x)H(y)$$

Seja ρ_n o menor valor satisfazendo a desigualdade:

$$P\{D > \rho_n | F(x, y) = G(x)H(y)\}$$

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.

4.6 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.

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

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 **x** formada por a partes (**x**-partição) e uma partição de D nos valores de **y** composta por b partes (**y**-partição). Chamaremos a intersecção entre uma parte da **x**-partição e uma parte da **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.

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*.

4.7 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 **x** e **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*.

4.8 Medida de Heller, Heller e Gorfine (HHG)

Heller, Heller e Gorfine propõem um teste de independência que tenha elevado poder estatístico e que seja consistente.

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

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:

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

4.9 Simulações

Nesta seção serão descritas todas as simulações realizadas para o estudo comparativo.

4.10 Aplicações em dados de expressão gênica

4.10.1 Dados utilizados no presente trabalho

Nessa seção farei uma breve descrição dos dados utilizados.

4.10.2 Aplicação das medidas estudas aos dados de expressão gêncica

Nessa seção explicarei o processamento dos dados biológicos e descreverei a aplicação realizada.

Capítulo 5
Resultados

Nessa seção, serão exibidos os resultados do trabalho.

Capítulo 6
Discussão

Nesta seção, discutirei os resultados obtidos.

Capítulo 7
Conclusões

Conclusões e considerações finais sobre o trabalho.

Referências Bibliográficas

[1] Fujita A, Sato JR, Demasi MA, Sogayar MC, Ferreira CE, and 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.

[2] Reshef DN, Reshef YA, Finucane HK, Grossman SR, McVean G, Turnbaugh PJ, Mitzenmacher M Lander ES, and Sabeti PC. Detecting novel associations in large data sets. *Science*, 334(6062):1518–1524, 2011.

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

[4] Heller R, Heller Y, and Gorfine M. A consistent multivariate test association based on ranks of distances. Front for the Mathematics ArXiv, under Statistics, arXiv:1201.3522v1, 2012.

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

[6] 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.

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

[8] Wikipedia. Bootstrapping (statistics) — Wikipedia, the free encyclopedia, 2012. `http://en.wikipedia.org/wiki/Bootstrapping_(statistics)`.

[9] Wikipedia. Kendall tau rank correlation coefficient — Wikipedia, the free encyclopedia, 2012. `http://http://en.wikipedia.org/wiki/Kendall_tau_rank_correlation_coefficient`.

[10] Wikipedia. Mutual information — Wikipedia, the free encyclopedia, 2012. `http://en.wikipedia.org/wiki/Mutual_information`.

[11] Wikipedia. Pearson product-moment correlation coefficient — Wikipedia, the free encyclopedia, 2012. `http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient`.

[12] Wikipedia. Spearman's rank correlation coefficient — Wikipedia, the free encyclopedia, 2012. http://en.wikipedia.org/wiki/Spearman's_rank_correlation_coefficient.

Parte II

Parte Subjetiva

Desafios e frustrações

Nessa seção descreverei os principais desafios e frustraões que encontrei no desenvolvimento do trabalho.

Paralelo entre o trabalho de formatura e as disciplinas do BCC

Farei um breve comentário sobre as disciplinas que mais contribuiram para o meu trabalho de formatura.

Trabalhos futuros

Os possíveis passos que tomarei se continuar trabalhando na área do TCC.