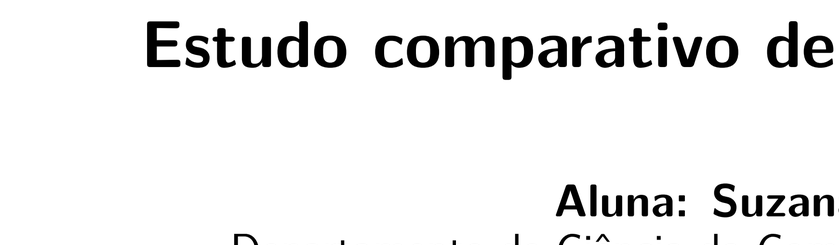
Estudo comparativo de medidas de dependência e aplicações em dados de expressão gênica

Aluna: Suzana de Siqueira Santos **Orientador: André Fujita**
Departamento de Ciência da Computação, Instituto de Matemática e Estatística, Universidade de São Paulo

Resumo

Diversas áreas do conhecimento utilizam-se de medidas de dependência estatísticas para detectar associações entre variáveis de um determinado conjunto de dados. Dada a grande diversidade de relações possíveis entre as variáveis sob análise (linear **(a)**, não-linear **(b)**, e as que nem sequer são funções **(c)**), é desejável a utilização de medidas que sejam capazes de reconhecer numerosos tipos de associação. Muitas relações interessantes entre os dados deixam de ser percebidas por métodos tradicionais como Pearson, Spearman e Kendall. No presente trabalho, comparamos alguns métodos clássicos utilizados para detectar associações e outros desenvolvidos mais recentemente, e aplicamos os mesmos em dados biológicos reais.



(a) Linear (b) Quadrática (c) Circunferência

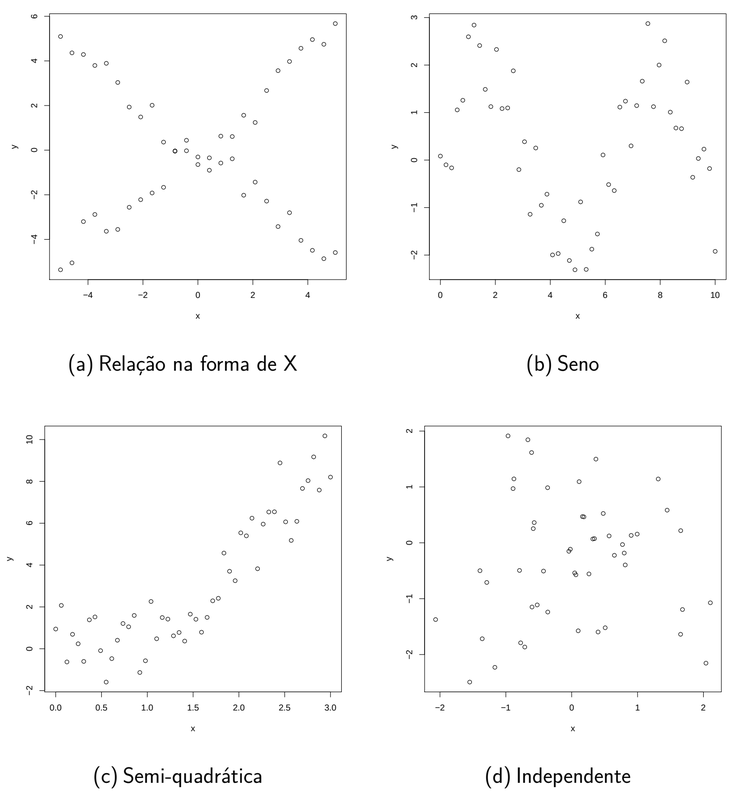
Figura: 1. Exemplo de tipos de associações entre dados

Objetivos

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 [1] e Spearman [1], o tau de Kendall [6], a informação mútua (IM) [7], a medida D de Hoeffiding [5], o Coeficiente de Informação Máxima (CIM) [2], a correlação de distância (Dcor) [3] e a medida de Heller, Heller e Gorfine (HHG) [4], com posterior aplicação dados biológicos de expressão gênica advindos de tecnologia de microarranjos de DNA

Simulações

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



(a) Relação na forma de X (b) Seno

(c) Semi-quadrática (d) Independente

Figura: 2. Figura ilustrativa das simulações

Mais informações

Mais informações sobre o trabalho no sítio:
http://www.linux.ime.usp.br/~suzanasantos/mac499

Contato: suzana@vision.ime.usp.br

Agradecimentos

Agências financiadoras: Conselho Nacional de Desenvolvimento Científico e Tecnológio (**CNPq**) e Fundação de Amparo à Pesquisa do Estado de São Paulo (**FAPESP**).

Testes de independência

Segue uma breve exposição dos 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.

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

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

onde o coeficiente de correlação de Pearson r_p é dado por $r_p = \frac{n\Sigma x_i y_i - \Sigma x_i \Sigma y_i}{\sqrt{n\Sigma x_i^2 - (\Sigma x_i)^2}\sqrt{n\Sigma y_i^2 - (\Sigma y_i)^2}}$

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

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

onde r_s é a aplicação do coeficiente de Pearson nos dados convertidos em postos.

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

Informação Mútua (IM): 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. O teste de independência é construído com com base na técnica computacional de *bootstrap*.

Medida D de Hoeffding: Rejeitamos H_0 se e somente se $30D > \rho_n$, onde:

$$\rho_n = \sqrt{\frac{2(n^2+5n-32)}{9n(n-1)(n-3)(n-4)\alpha}},$$
$$D = \frac{(n-2)(n-3)D_1 + D_2 - 2(n-2)D_3}{n(n-1)(n-2)(n-3)(n-4)},\ D_1 = \Sigma_{i=1}^n Q_i(Q_i-1),$$
$$D_2 = \Sigma_{i=1}^n (R_i-1)(R_i-2)(S_i-1)(S_i-2),$$
$$D_3 = \Sigma_{i=1}^n (R_i-2)(S_i-2)Q_i.$$

α é o nível de significância do teste, R_i é o posto de x_i, S_i é o posto de y_i e Q_i é o número de pontos com ambos os valores de x e y menores do que o i-ésimo ponto.

Coeficiente de informação Máxima: O Coeficiente de Informação Máxima (CIM) é a maior informação mútua (normalizada entre 0 e 1) de todas as grades no gráfico xy dos valores das amostras de X e Y. O teste estatístico se baseia na técnica computacional de *bootstrap*.

Correlação de distância (Dcor): A correlação de distância entre duas variáveis aleatórias X e Y é dada por:

$$dCor(X,Y) = \frac{dCov(X,Y)}{\sqrt{dVar(X)dVar(Y)}}$$

Onde $Dcov$ e $Dvar$ são calculados empiricamente a partir das matrizes de distâncias de x e y. O teste estatístico é implementado como um teste de permutação.

Medida de HHG: O teste de independência para a medida de HHG é baseado nas distâncias dos pares entre os valores de X e os valores de Y, respectivamente, $\{d_X(x_i, x_j) : i, j \in \{1, ..., n\}\}$ e $\{d_Y(y_i, y_j) : i, j \in \{1, ..., n\}\}$.
A estatística de teste é uma função dos postos dessas distâncias.

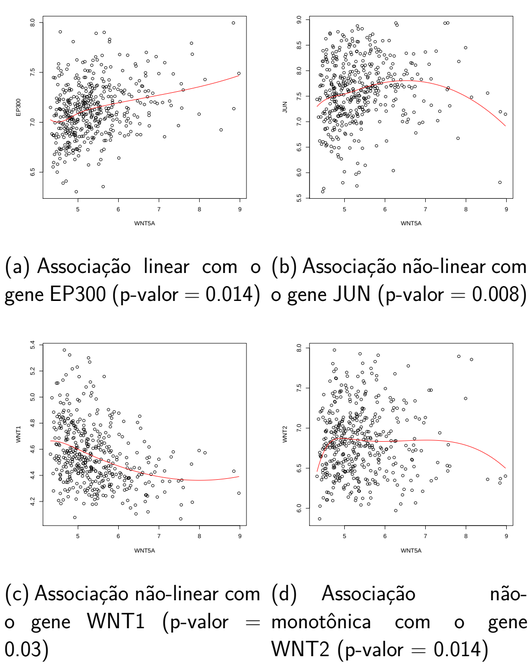
Resultados

Tabela: 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	50	0,51	0,50	0,50	0,51	0,57	0,51	0,52	0,58
	10	0,50	0,50	0,48	0,45	0,50	0,50	0,71	0,43
Independente com outliers	50	0,71	1,00	0,54	0,53	0,60	0,95	0,67	0,60
	10	0,87	0,89	0,56	0,52	0,48	0,57	0,26	0,41
Linear	50	1,00	1,00	1,00	1,00	1,00	1,00	1,00	1,00
	10	1,00	1,00	1,00	1,00	1,00	0,99	0,91	0,94
Linear com outliers	50	0,72	1,00	1,00	1,00	1,00	1,00	0,69	0,99
	10	0,86	0,95	0,72	0,75	0,78	0,86	0,28	0,70
Quadratica	50	0,21	1,00	0,18	0,23	1,00	1,00	1,00	1,00
	10	0,16	0,71	0,16	0,14	0,90	0,97	0,84	0,61
Quadratica com outliers	50	0,05	0,64	0,31	0,32	0,99	1,00	0,70	1,00
	10	0,14	0,16	0,28	0,23	0,70	0,78	0,12	0,43
Cubica	50	0,72	0,97	0,77	0,78	0,98	0,99	0,93	0,99
	10	0,34	0,45	0,32	0,28	0,43	0,54	0,75	0,35
Seno	50	0,40	0,98	0,42	0,42	0,99	1,00	1,00	1,00
	10	0,28	0,29	0,32	0,24	0,51	0,34	0,74	0,25
X	50	0,12	0,77	0,11	0,11	0,85	1,00	1,00	0,99
	10	0,09	0,03	0,14	0,11	0,00	0,66	0,85	0,06
Circunferencia	50	0,09	0,38	0,15	0,18	0,95	0,99	0,97	0,97
	10	0,09	0,10	0,24	0,20	0,64	0,71	0,75	0,17

Tabela: 2. 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
Robusta à presença de outliers			X	X	X	X		X



(a) Associação linear com o gene EP300 (p-valor = 0.014) (b) Associação não-linear com o gene JUN (p-valor = 0.008)

(c) Associação não-linear com o gene WNT1 (p-valor = 0.03) (d) Associação não-monotônica com o gene WNT2 (p-valor = 0.014)

Figura: 3. Exemplos de associações encontradas com o gene WNT5A nos dados de expressão gênica (em escala logarítmica) de 441 amostras de adenocarcinoma

Referências

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

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

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