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

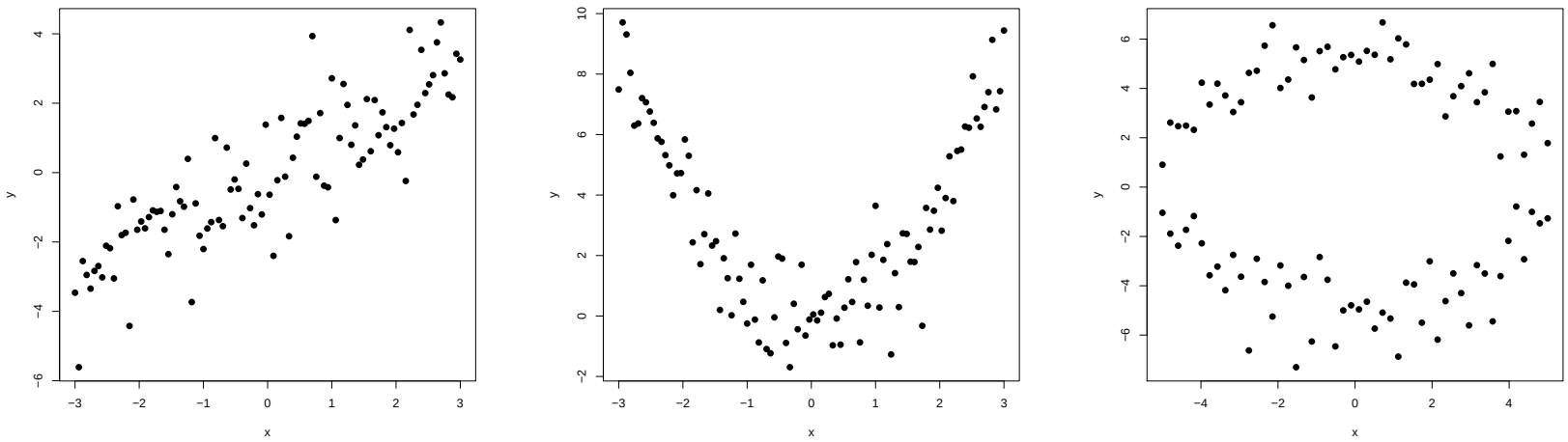
Trabalho de Conclusão de Curso

Aluna: Suzana de Siqueira Santos
Orientador: André Fujita

Novembro de 2012

O que são medidas de dependência?

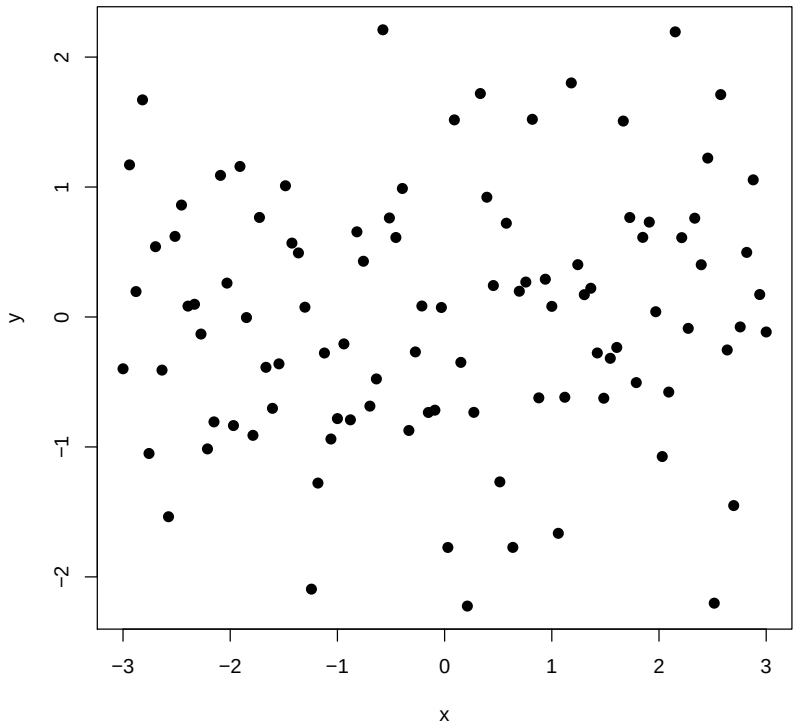
Quando duas variáveis de um conjunto de dados são dependentes, esperamos existir uma associação entre os valores observados:



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

Figura: Exemplo de tipos de associações entre dados

Já, quando duas variáveis são independentes, não esperamos identificar associações entre os valores observados:



Dependência entre duas variáveis aleatórias

Formalmente, duas variáveis aleatórias, X e Y, com funções de distribuição acumulada $P_X(x)$ e $P_Y(y)$, respectivamente, são dependentes se:

$$P_{XY}(x, y) \neq P_X(x)P_Y(y)$$

onde $P_{XY}(x, y)$ é a função de distrubuição acumulada conjunta de X e Y

O que são medidas de dependência?

Medidas de dependência quantificam associações entre variáveis aleatórias

Quando utilizamos medidas de dependência?

Exemplos

- O peso das mães está relacionado com o peso dos filhos?
- O desempenho escolar está relacionado com a quantidade de horas de estudo semanal?
- A quantidade de filhos tem associação com renda familiar?
- Dependência entre áreas do cérebro

Por que comparar medidas de dependência?

- Existem diversas medidas na literatura
- Algumas medidas são bastante recentes: medida de Heller, Heller e Gorfine (ainda não publicada), Coeficiente de Informação Máxima (2011), Correlação de Distância (2007)
- Não há muitos estudos comparando tais medidas com outras mais tradicionais

Como realizamos o estudo comparativo?

Escolhemos as seguintes medidas para o estudo:

- Correlação de Pearson
- Correlação de Distância (Dcor)
- Correlação de Spearman
- Tau de Kendall
- Medida D de Hoeffding
- Medida de Heller, Heller e Gorfine (HHG)
- Informação Mútua (IM)
- Coeficiente de Informação Máxima (CIM)

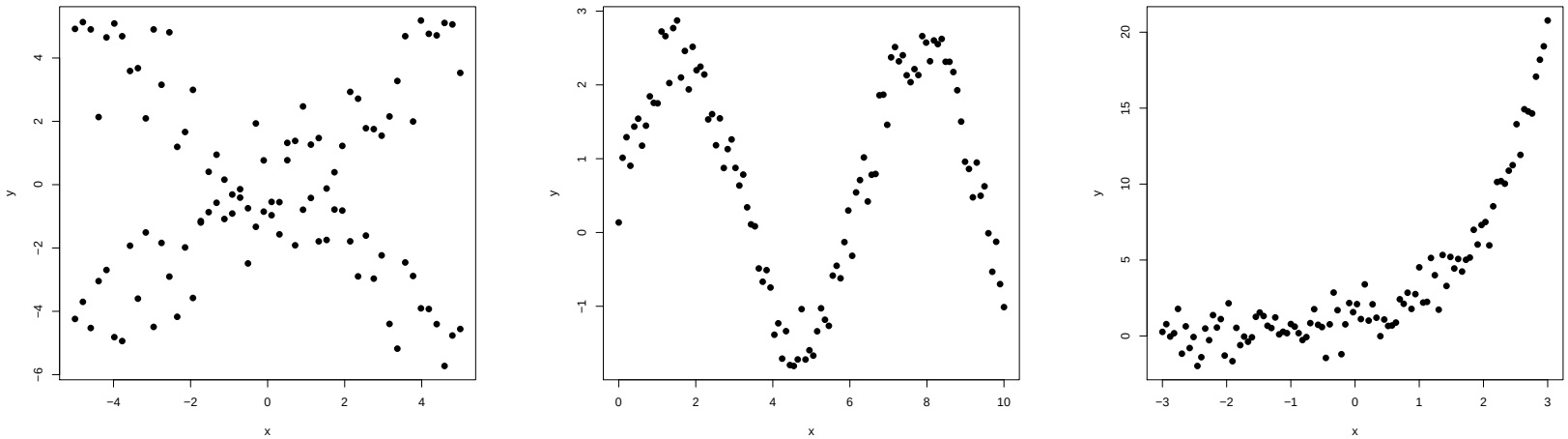
Como realizamos o estudo comparativo?

Para cada medida, realizamos testes de independência em diversos tipos de dados gerados com a ferramenta R.

O teste estatístico realizado tem a seguinte descrição:
$H_0 : X$ e Y são independentes
$H_1 : X$ e Y não são independentes

Simulações

Simulamos diversos tipos de dados, variando o tamanho das amostras:

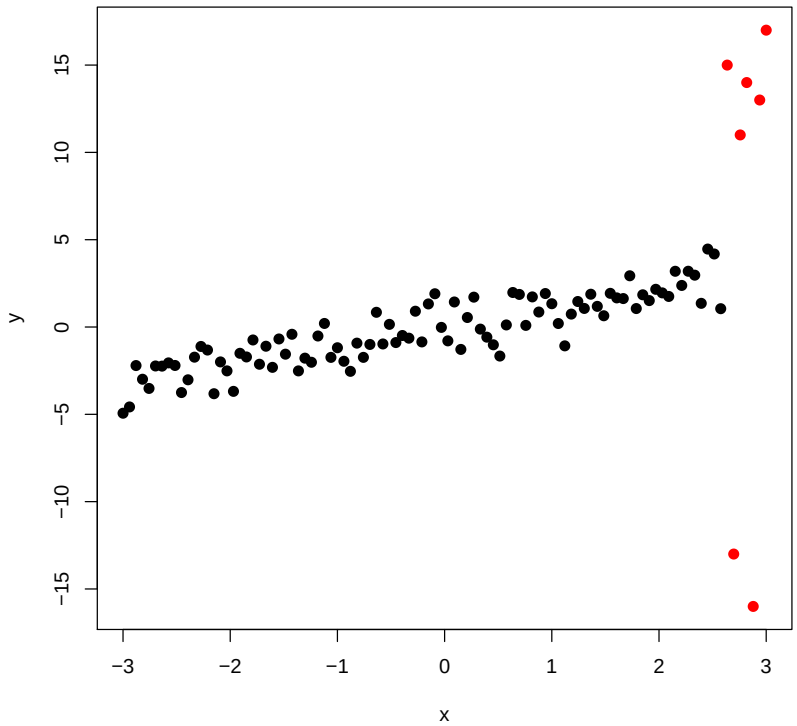


(a) Não funcional (b) Não monotônica (c) Monotônica

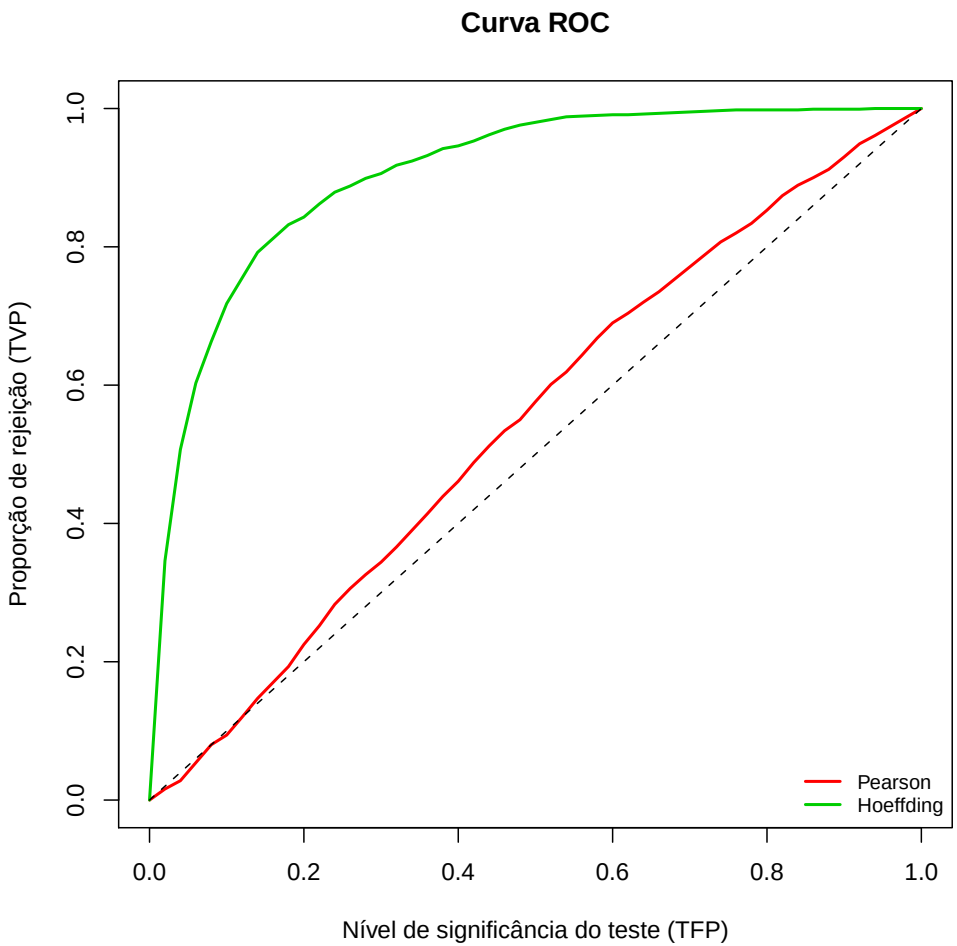
Figura: Ilustração das simulações

Simulações

Inserimos outliers em algumas amostras para observar os efeitos nos testes de independência:



Comparação das medidas



Resultados das simulações

Tabela: 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
Quadrática	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
Quadrática 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
Cúbica	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
Circunferência	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

Síntese dos resultados

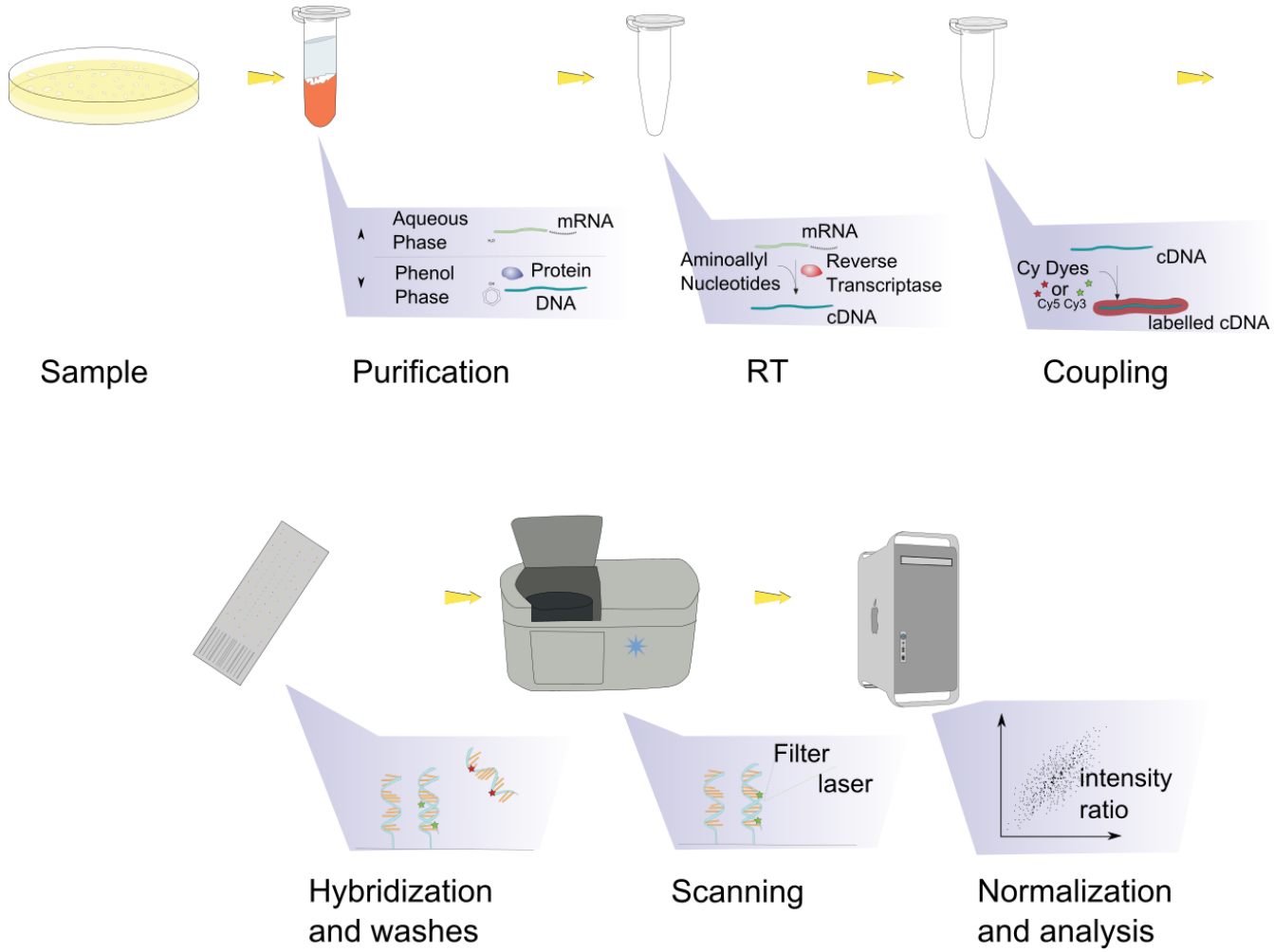
Tabela: Tipo de associações detectadas por cada medida

Medida	Linear	Monotônica não linear	Não-monotônica	Robusta à presença de outliers
Pearson	X			
Dcor	X	X	X	
Spearman	X	X		X
Kendall	X	X		X
Hoeffding	X	X	X	X
HHG	X	X	X	X
IM	X	X	X	
CIM	X	X	X	X

Como aplicamos os resultados obtidos nos dados de expressão gênica?

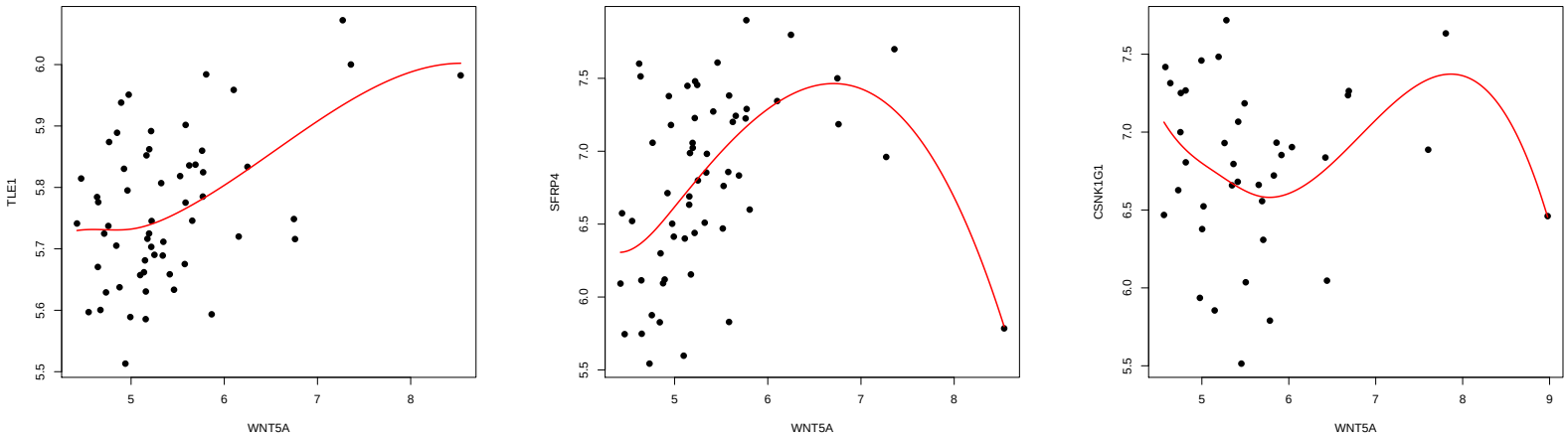
- É difícil estimar a sobrevida de um paciente no estágio I do câncer de pulmão
- Com o auxílio da pesquisadora Asuka Nakata do Instituto de Ciências Médicas da Universidade de Tóquio, selecionamos genes conhecidos pela literatura para estudar a dependência com o gene WNT5A, nos dados de expressão gênica de tumores de pulmão
- ~400 amostras de adenocarcinomas
- ~200 amostras de tumores no estágio I

Microarranjos de DNA



Fonte: `http://en.wikipedia.org/wiki/File:Microarray_exp_horizontal.svg`

Testes de independência com o gene WNT5A



(a) Associação linear com o gene TLE1 (p-valor = 0.025)

(b) Associação não linear com o gene SERP4 (p-valor = 0.029)

(c) Associação não-monotônica com o gene CSNK1G1 (p-valor = 0.018)

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

Agradecimentos

Agradecimentos

- Apoio financeiro: PIBIC/CNPq

Agradecimentos

- Apoio financeiro: PIBIC/CNPq
- Orientação do professor André Fujita


NÃO SE ESQUEÇA
INAUGURAÇÃO EM BREVE


Voluntários
Telefônica
Brasil

Obrigada a todos pela atenção e paciência!