A regra dos três números para o cálculo de uma medida de correlação robusta

Gustavo H. Esteves * Diana Maia*

Abril de 2010

Resumo

Um dos problemas mais comuns na Estatística é o cálculo de uma medida de correlação robusta, isto é, uma medida que não seja influenciada por pontos discrepantes (*outliers*) presentes no conjunto de dados. Neste trabalho é apresentado um método, baseado na técnica de *leave one out* da teoria de discriminadores lineares, que ataca este problema e define uma regra, chamada aqui de *regra dos três números*, que usa a informação do mínimo, da média (ou mediana) e do máximo entre n valores de correlação linear de Pearson, onde n é o número de observações da amostra, para estimar um valor de correlação robusto.

O coeficiente de correlação é uma medida numérica entre duas variáveis aleatórias contínuas, que foi introduzida inicialmente por Karl Pearson no início do século XIX e mede a relação (ou associação) linear entre as duas variáveis. Atualmente esta medida é muito utilizada na maioria das áreas de pesquisa, tais como as Ciências Biológicas, Sociais, Políticas ou Econômicas.

A área de Biologia é uma das áreas da ciência que tem apresentado um trabalho muito intenso, envolvendo a área de Estatística, nos últimos anos (de la Funte *et al.*, 2004; Steuer *et al.*, 2003). Com o advento de técnicas relativamente novas que podem medir os níveis de expressão de milhares de genes simultaneamente, como SAGE (Velculescu *et al.*, 1995) e, especialmente, a tecnologia de *microarray* (DeRisi *et al.*, 1996; Schena *et al.*, 1995), medidas

*Departamento de Estatística - Centro de Ciências e Tecnologia - Universidade Estadual da Paraíba, Campina Grande - PB, Brazil.

de associação, tais como a correlação linear de Pearson se tornaram ferramentas importantes para a busca de pares de genes que possam mostrar algum tipo de alteração nos seus padrões de associação. Isso é particularmente interessante para a construção de esboços de redes genéticas, como a construção de redes de relevância (Butte *et al.*, 2000; Butte & Kohane, 2003). Recentemente, também usamos técnicas deste tipo para inferir possíveis redes gênicas em dois trabalhos realizados, um deles em conjunto com um grupo de pesquisa do Hospital do Câncer A. C. Camargo de São Paulo-SP (Esteves *et al.*, 2007; Gomes *et al.*, 2005).

Porém, a presença de pontos discrepantes (*outliers*) – até mesmo um único valor deste tipo – pode interferir drasticamente no cálculo do coeficiente de correlação. Existem alguns métodos estatísticos para lidar com este problema, tais como Abdullah (1990), Gnanadesikan & Kettenring (1972), Hampel *et al.* (1986), Huber (1981) e Lawrence & Arthur (1990), que tentam remover os *outliers* previamente ao cálculo do coeficiente ou durante este cálculo. Infelizmente, em geral, estes métodos são de difícil interpretação ou impõem limitações na implementação do método.

Para tentar contornar estes problemas, este trabalho apresenta uma estratégia que usa um método do tipo *leave-one-out*, muito comum em técnicas de análise multivariada para análise de discriminação, que calcula uma medida robusta para a correlação linear de Pearson. Neste método, suponha que tem-se dois vetores n-dimensionais e deseja-se calcular um coeficiente de correlação robusto entre esses dois vetores de observações. Então, uma observação é removida a cada etapa e o coeficiente de correlação linear de Pearson é calculado com as $n-1$ observações restantes, o processo é repetido até que todas as observações tenham sido removidas gerando, assim, um total de n valores de correlação diferentes. A idéia é usar estes valores de correlação para calcular enfim uma medida robusta para a correlação linear de Pearson. Note que nesta abordagem, qualquer outra medida de associação pode ser usada.

Matematicamente, dadas duas variáveis aleatórias quantitativas X e Y, a correlação linear de Pearson entre essas duas variáveis é definida como

$$\rho(X,Y) = \frac{\mathbb{E}(XY) - \mathbb{E}(X)\mathbb{E}(Y)}{\mathrm{DP}(X)\mathrm{DP}(Y)}, \tag{1}$$

onde $\mathbb{E}$ e DP denotam a esperança e o desvio padrão das variáveis. Dadas $\mathbf{x}$ e $\mathbf{y}$ duas amostras aleatórias das variáveis X e Y, com tamanho n, a correlação linear de Pearson

pode ser estimada por

$$r(\mathbf{x},\mathbf{y}) = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{(x_i-\overline{x})^2}\sqrt{(y_i-\overline{y})^2}}, \tag{2}$$

onde $\overline{x}$ e $\overline{y}$ são as médias amostrais de $\mathbf{x}$ e $\mathbf{y}$. É importante enfatizar que a correlação linear de Pearson mede o grau de associação linear entre duas variáveis, que varia no intervalo $[-1,1]$, onde -1 representa forte associação negativa, 1 representa forte associação positiva e zero representa ausência total de associação linear.

Um problema bastante comum em estatística com a utilização do valor de correlação linear dado pela equação (2) é a influência de valores extremos (*outliers*); muitas vezes encontramos valores de correlação extremamente altos, ou extremamente baixos, que se devem simplesmente a presença de um ponto muito discrepante em relação aos demais. Para contornar este tipo de problema, definimos uma medida robusta onde calcula-se a correlação entre $\mathbf{x}$ e $\mathbf{y}$, ambos com n elementos, através de um processo iterativo onde é removida uma observação por vez e é calculado o valor de correlação linear de Pearson entre as $n-1$ observações restantes. No final desta etapa, temos um vetor $\mathbf{r}$ com n valores de correlações diferentes. Se o conjunto de dados não apresenta valores extremos todos os n valores calculados estão muito próximos e podemos tomar qualquer um deles como estimativa sem maiores problemas. Entretanto, quando o conjunto de dados apresenta algum valor extremo temos $n-1$ valores de correlação "contaminados" e apenas um valor que não sofre a influência deste ponto. Assim, estabelecemos a seguinte regra de decisão para definir um valor robusto para r denotado por r'

$$r'(\mathbf{x},\mathbf{y}) = \begin{cases} \min(\mathbf{r}), & \text{se} \quad \max(\mathbf{r}) - \overline{r} < \overline{r} - \min(\mathbf{r}), \\ \max(\mathbf{r}), & \text{se} \quad \max(\mathbf{r}) - \overline{r} > \overline{r} - \min(\mathbf{r}), \end{cases} \tag{3}$$

onde $\min(\mathbf{r})$, $\max(\mathbf{r})$ e $\overline{r}$ são os valores mínimo, máximo e médio dos valores de correlação calculados. Ou seja, decide-se pelo valor mínimo ou máximo, de acordo com o que esteja mais distante da média; também é possível usar o valor mediano aqui. Note que esta medida de correlação robusta é similar à correlação *jackknife* (Heyer *et al.*, 1999), onde o mesmo procedimento iterativo é usado, mas com o critério de decisão dado por $r'(\mathbf{x},\mathbf{y}) = \min(\mathbf{r})$. Entretanto, a correlação robusta dada pela equação (3) consegue lidar com situações onde temos uma forte dependência linear viesada pela presença de um ponto discrepante, o que não é verdade para a correlação *jackknife*.

Quando o interesse é estimar a significância dos valores de correlação robusta obtidos, é possível utilizar estratégias de permutação dos dados (técnicas de *bootstrap*). Nos métodos

de permutação, os valores observados na amostra são permutados independentemente um número grande, digamos B, de vezes e os coeficientes de correlação robusta são recalculados em cada repetição do processo. Assim, podemos contar o número de vezes, digamos b, em que se obtém valores maiores que o valor observado nos dados originais e definimos o nível descritivo do teste como b/B. Se o interesse for o teste bicaudal, devemos tomar os valores absolutos das estatísticas, se o teste for feito à esquerda ou direita devemos contar o número de vezes em que a estatística permutada é menor ou maior que o valor originalmente observado. Esta estratégia é especialmente interessante por não fazer nenhum tipo de suposição sobre a distribuição dos dados.

Referências

Abdullah, M. B. (1990). On a robust correlation coefficient. *The Statistician* **39**, 455–460.

Butte, A. J. & Kohane, I. S. (2003). *The analysis of gene expression data*, cap. Relevance networks: a first step towards finding genetic regulatory networks within microarray data. New York: Springer Verlag.

Butte, A. J., Tamayo, P., Slonim, D., Golub, T. R. & Kohane, I. S. (2000). Discovering functional relationships between RNA expression and chemotherapeutic susceptibility using relevance networks. *PNAS* **97**(22), 12182–12186.

de la Funte, A., Bing, N., Hoeschele, I. & Mendes, P. (2004). Discovery of meaningful associations in genomic data using partial correlation coefficients. *Bioinformatics* **20**(18), 3565–3574.

DeRisi, J., Penland, L., Brown, P. O., Bittner, M. L., Meltzer, P. S., Ray, M., Chen, Y., Su, Y. A. & Trent, J. M. (1996). Use of a cDNA microarray to analyse gene expression patterns in human cancer. *Nat Genetics* **14**, 457–460.

Esteves, G. H., Simoes, A. C. Q., Souza, E., Dias, R. A., Ospina, R. & Venancio, T. M. (2007). New insights about host response to smallpox using microarray data. *BMC Syst Biol* **1**, 38. URL `http://dx.doi.org/10.1186/1752-0509-1-38`.

Gnanadesikan, R. & Kettenring, J. R. (1972). Robust estimates, residuals, and outlier detection with multiresponse data. *Biometrics* **28**(1), 81–124.

Gomes, L. I., Esteves, G. H., Carvalho, A. F., Cristo, E. B., Hirata, R., Martins, W. K., Marques, S. M., Camargo, L. P., Brentani, H., Pelosof, A., Zitron, C., Sallum, R. A., Montagnini, A., Soares, F. A., Neves, E. J. & Reis, L. F. L. (2005). Expression profile of malignant and nonmalignant lesions of esophagus and stomach: differential activity of functional modules related to inflammation and lipid metabolism. *Cancer Res* **65**(16), 7127–7136. URL `http://dx.doi.org/10.1158/0008-5472.CAN-05-1035`.

Hampel, F. R., Ronchetti, E. M., Rousseeuw, P. J. & Stahel, W. A. (1986). *Robust statistics: the approach based on influence functions*. John Wiley & sons.

Heyer, L. J., Kruglyak, S. & Yooseph, S. (1999). Exploring expression data: identification and analysis of coexpressed genes. *Genome Research* **9**, 1106–1115.

Huber, P. J. (1981). *Robust Statistics*. John Wiley & sons.

Lawrence, K. D. & Arthur, J. L. (eds.) (1990). *Robust Regression Analysis and Applications*. New York: Marcel Dekker, Inc.

Schena, M., Shalon, D., Davis, R. W. & Brown, P. O. (1995). Quantitative monitoring of gene expression patterns with a complementary DNA microarray. *Science* **270**, 467–470.

Steuer, R., Kurths, J. & Weckwerth, W. (2003). Observing and interpreting correlations in metabolic networks. *Bioinformatics* **19**(8), 1019–1026.

Velculescu, V. E., Zhang, L., Vogelstein, B. & Kinzler, K. W. (1995). Serial analysis of gene expression. *Science* **270**, 484–487.