Detecting a Wide Diversity of Associations in Very Large Data Sets

Christopher Pardy[1] and Susan R Wilson[1,2,]
[1] University of New South Wales, Sydney, Australia
[2] Australian National University, Canberra, Australia
[3] Corresponding author: Sue Wilson, e-mail: Sue.Wilson@anu.edu.au

Abstract

Major challenges arising from today's "data deluge" include how to handle the commonly occurring situation of different types of variables (say, continuous and categorical) being simultaneously measured, as well as how to assess accompanying flood of questions. Based on information theory, a bias-corrected mutual information (BCMI) measure of association that is valid and estimable between all basic types of variables has been proposed. It has the advantage of being able to identify non-linear as well as linear relationships. Based on the BCMI measure a novel exploratory approach to finding associations in large data sets has been developed. These associations can be used as a basis for finding clusters and networks, for example, in large data sets in which different types of variables have been collected on each individual (or unit). The application of this exploratory approach is very general. Comparisons will be made with the recently proposed measure, maximal information coefficient (MIC). Illustrative examples include exploring relationships (i) between social, economic, health and political indicators from the World Health Organisation and partners, and (ii) in genomic (say, gene expression and genotypic) and clinical data.

Keywords: mutual information, high dimensional data, exploratory data analysis, non-linear relationships identification

1. Introduction

In many scientific fields, massive data sets are being produced by modern, and continually evolving, technologies. For instance, it is now feasible to relatively routinely obtain many tens of thousands of measurements on a single individual. Due to the current cost of these technologies, the number of individuals usually is relatively small. Analyses of such data are posing substantial computational and statistical challenges that include how to simultaneously analyse different types of data. For example, in genomics the types of data collected for each individual may included microarray (continuous) data, genotypic (discrete, categorical) data, as well as other measurements of gene expression (RNA-Seq data), epigenomic (such as DNA methylation) data, clinical information, and so on. Often a general goal is to infer (biochemical) networks from such data. There is a lack of a single exploratory measure for all types of data for situations where no single variable is a sole primary outcome of interest. Currently, most methods to deal simultaneously with such data involve either making the continuous variables discrete (resulting in a loss of information) or making discrete variables continuous; neither approach is particularly satisfactory.

Mutual Information (MI) has been widely used for finding non-linear relationships, in particular for discrete data comparisons and for continuous data comparisons. For 'mixed' comparisons, namely between a continuous and a discrete variable, MI was initially deemed to pose too many computational problems for automatic application (Dawy et al, 2006). We have developed a nonparametric approach, bias-corrected mutual information (BCMI), which can be successfully automated. In the following this approach is briefly outlined, and compared with the Maximal Information Coefficient (MIC) proposed by Reshef et al (2011) that also is based on MI. Application of both methods is made to some World Health Organisation data, and to data from part of a moderate-sized (genomic) data set.

2. Outline of the Background Theory

For two discrete random variables X and Y with joint probability mass function given by $P(X = x, Y = y) = p(x, y)$ MI is defined by $I(X,Y) = \sum_x \sum_y p(x,y) \log(\frac{p(x,y)}{p(x)p(y)})$.

For continuous variable, sums are replaced by integrals (Cover and Thomas, 2006). Where one of the variables (X) is discrete, the other (Y) continuous, MI can be shown to be $I(X,Y) = \sum_i p_i \int_y f_i(y) \log(\frac{f_i(y)}{f(y)}) dy$ where $f_i(y)$ represents the conditional density of $Y|X = x_i$ and $f(y) = \sum_i p_i f_i(y)$; this result also can be deduced from Dawy et al (2006).

To estimate MI between two continuous variables, we use non-parametric kernel density estimators (Wand and Jones, 1995). To estimate MI between a continuous variable and a discrete variable, what we will refer to as a mixed comparison, our preferred estimator is based on the asymptotically optimal Epanechnikov kernel; details in Pardy (2013).

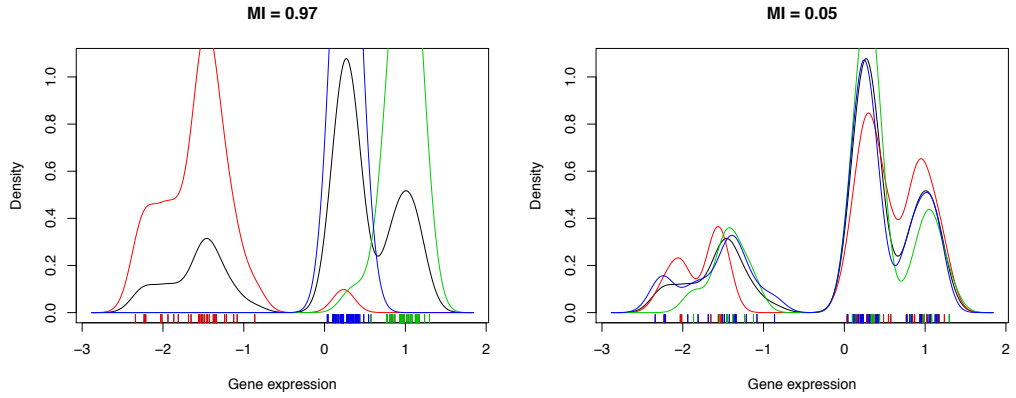

Figure 1. Kernel density estimates of Aqr gene expression according to rs3149884 genotype (liver tissue data; see text). The rug plot and correspondingly coloured densities show a clear separation of groups in the lhs plot compared to the rhs plot where the grouping variable (genotype) has been randomly permuted. The black curve is the combined density.

A visual example of a mixed comparison estimator is given in Figure 1. On the left, there are three estimated conditional distributions with little overlap and the MI score is high, 0.97, while the figure on the right, that was determined from a random permutation of group membership, shows substantial overlap of the three conditional distributions and the MI score is low, 0.05

Since the MI estimator was found to be biased, a jackknife-based correction has been developed that potentially may reduce the mean square error (MSE = bias^2+variance); see Efron and Gong, 1983. Further, the jackknife was chosen over the bootstrap because its deterministic nature makes it easy to re-use calculations (and thus make it faster), and moreover the number of resamples is limited by the sample size. The lower bound of MI is zero (corresponding to no association between the variables), while bias-corrected MI (BCMI) sometimes can have small negative values. Such values are not of interest in our exploratory data analysis to discover non-zero relationships between variables. Extensive simulations have been undertaken to both evaluate the bias correction and to choose the smoothing bandwidth parameter needed for our estimation procedure. Overall the simulations showed the following: (i)

estimation bias is a substantial issue that can be addressed reasonably well using the jackknife bias correction, and (ii) that taking plug-in level 3 generally works well for the smoothing bandwidth (in contrast to the value 2 that is the usual default).

Further, it is noted that a novel Kolmogorov-Smirnov test approach has been developed to determine p-values for evaluation of whether the cumulative distributions are the same across categories; details not shown. Simulations have shown the nominal level to be quite conservative.

A software package for the R statistical environment has been developed and is available at http://r-forge.r-project.org/projects/mpmi/. The software is fast and easily parallelised with computational kernels written in Fortran for speed and portability.

In Reshef et al (2011), MIC is proposed as a nonparametric association measure designed for comparisons between variables of any type. In summary, the MIC association value is based on searching through an increasingly fine 'ragged' grid of discretisations of a scatterplot to find the partitioning that results in the greatest MI value. MI is calculated between the discrete variables induced by each partition with the final MIC being a function of the maximum obtained MI.

3. Results

First we consider part of the social, economic, health and political indicators data from the World Health Organisation global health data that were analysed by Reshef et al (2011), in particular comparing: A - ' income per person' with 'prevalence of female obesity'; B - 'health expenditure per person' with 'under 5 mortality rate' and C- 'cardiovascular mortality' with 'life expectancy'; the respective plots are given in Figure 2. These relationships were all determined by MIC to be significant.

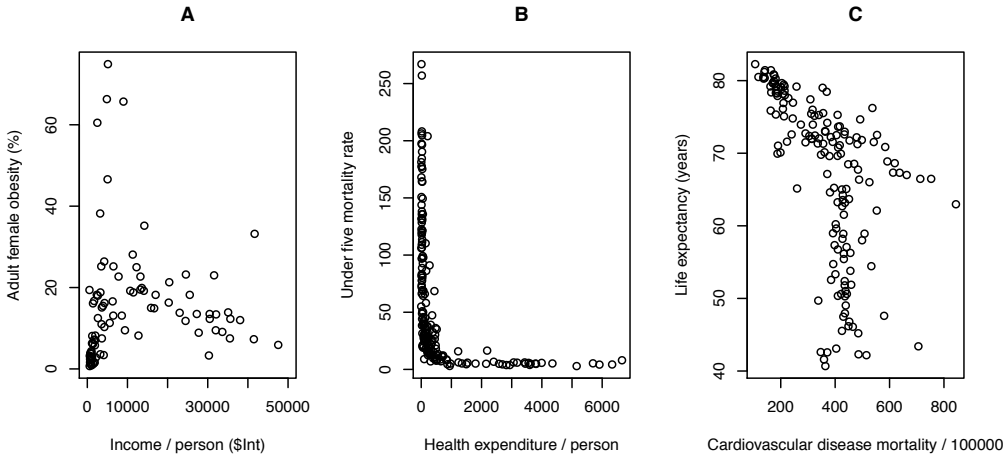

Figure 2. Plots of some interesting nonlinear relationships in the WHO data.

For comparison A, that is an example of the superposition of a linear and a quadratic relationship (and that according to Reshef et al is distinguishing the Pacific Islands, where female obesity is a sign of status from the rest of the world), the MI is 0.52 that drops substantially to 0.37 after jackknife bias correction. For comparison B, that is an example of a L-shaped relationship, MI is 0.69, BCMI is 0.63. For comparison C, the shape is unusual, possibly indicating subgroups in the data, and MI is 0.58, BCMI is 0.52. Our approach also shows these relationships are significant, but is 18 times faster than MIC. This is important to keep in mind when carrying out millions of comparisons as is becoming commonplace in modern genomic experiments.

Next we consider the experimental data that motivated this research, an F2 intercross dataset containing 1065 Single Nucleotide Polymorphisms (SNPs), each of which gives a 3–level (categorical) genotype, and 3421 gene expression levels in liver tissue

from 135 female mice, and available at
http://labs.genetics.ucla.edu/horvath/CoexpressionNetwork/MouseWeight/
The data underpinning Figure 1 are from this experiment.

Since the most novel aspect of our development of BCMI is to mixed comparisons, here we focus on these. Our exploratory approach is highly conservative. For the 3,664,665 'mixed' comparisons using the conservative Kolmogorov-Smirnov test, and Bonferroni adjustment for an overall type 1 error rate of 0.01, 890 relationships were identified. These relationships included non-monotonic associations, as well as associations across chromosomes that are of biological interest and not previously determined; details in Pardy (2013).

For exploratory data analysis, the identified BCMI values of interest can be imported into clustering routines and network analyses. Here we give an example using the open source software Cytoscape (Smoot et al, 2011). A representative network (from 15 that were found using the above criterion) is given in Figure 3. Nodes corresponding to gene expression values, and that we refer to as genes, are given red borders, while nodes corresponding to SNPs are coloured blue. Having determined the gene-SNP network, associations between genes were incorporated and indicated here with thicker lines; details not shown. Interestingly, the annotations indicate that these three genes are associated with cardiovascular function or disease. There are many SNP-SNP associations due to linkage disequilibrium arising from the F2 design, and to avoid clutter these are not depicted in our illustrative network. It is noted that interactive networks for the experimenters to explore can be obtained based on this approach.

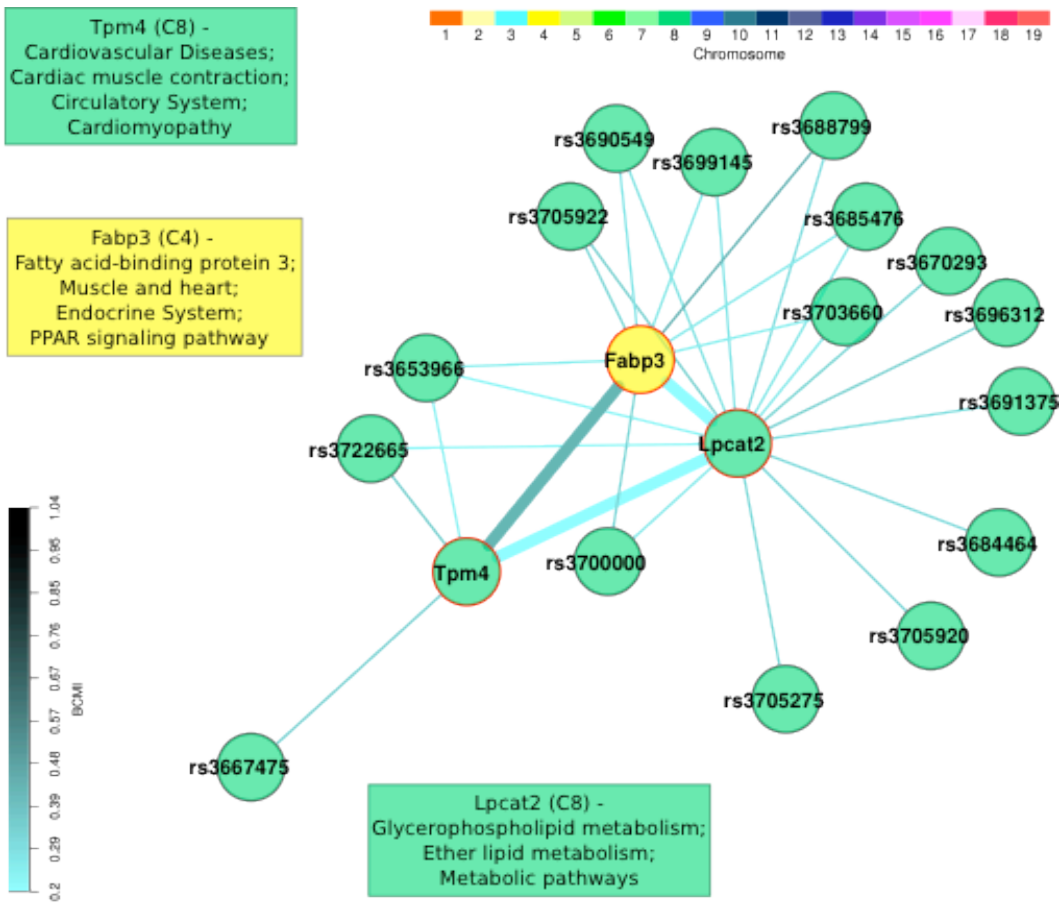

Figure 3. Network plot where green nodes are for chromosome 8, yellow nodes for chromosome 4, and known Kegg annotations are given; see text for further details.

Next, for these data, we found that there were many comparisons where MIC produced high values, while BCMI gave very low values, yet the plots gave no evidence of association. An example is given in Figure 4. The MIC value of 0.42 has a corresponding p-value of 0.0001 while the BCMI value of -0.03 indicates no association. The converse was not true, namely high BCMI values corresponded to high MIC values. As well, we applied the jackknife correction to MIC that reduced the value to 0.35. Further, we found that for MIC, swapping the labels of the heterozygous group H, with the homozygote group B changed the MIC value to 0.29, as MIC requires assignment of a numerical coding to categories. This is worrying, as although the allocation of 0, 1 and 2 can make sense for genotypic data (as the heterozyote is in some sense in-between the values of the 2 homozygotes), in many applications such an allocation of numerical values to categories is entirely arbitrary. No such assignment is needed for BCMI, so the value remains unchanged.

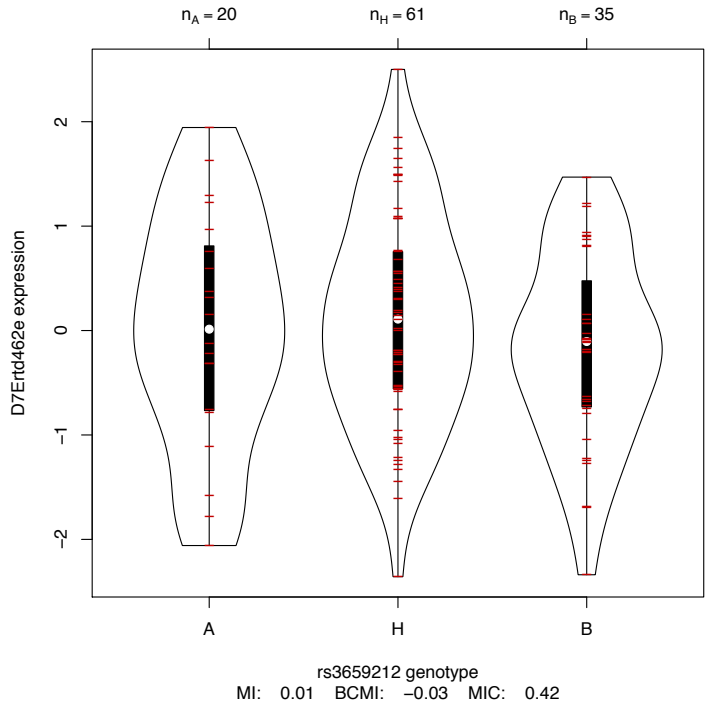

Figure 4. Box and violin plots of standardized D7Ertd462e gene expression grouped according to rs3659212 genotype. The data points are shown by red dashes. n_A, n_B and n_H are the sample sizes for the two homozygotes (AA and BB) and the heterozygote (AB) respectively (denoted A,B and H above).

4. Discussion and Conclusion

The motivation for this research was to find a useful approach for exploring the dependency structure of large and complex data sets. Using MI as a consistent framework for quantifying dependency, we can search for strong associations for any type of variable, continuous or discrete. The main advantage of our approach is that it provides a single association measure that can detect a wide variety of types of association between any combination of discrete or categorical variables.

To guide decisions as to which relationships to include at the next stage of, say, either clustering or determining networks, we have been using the extremely conservative Bonferroni bound. Other alternatives are obviously possible, based for example on false discovery rate (FDR) as advocated by Efron (2010) and Storey and Tibshirani (2003).

Simon & Tibshirani (2011) have proposed the use of the Brownian distance correlation as

an alternative to MIC for continuous measurements. In particular, they simulated pairs of observations with various functional relationships and various levels of noise, and compared calculated values of the statistics under investigation to those calculated when the relationship is absent. We have reproduced their simulation results with the addition of results from our BCMI estimator. We found that when considering power, BCMI does at least as well as MIC and better in some instances (with the exception of high frequency sine wave associations), and sometimes did as well as, or better than, Brownian distance correlation. Generally, BCMI had simulated power in between that for MIC and distance correlation (for continuous measurements). Further, simulations showed that for comparisons between discrete variables, the MIC values were very slightly less than the corresponding BCMI values and these, in turn, were found to be slightly less than the true MI values). Finally, we have shown that BCMI is robust in the presence of outliers, while for continuous comparisons we found MIC was too influenced by outliers or by groups of outliers.

In summary, to enable data containing extremely large numbers of both discrete and continuous variables (say, tens of thousands of variables) to be analysed, we have developed a single exploratory measure that then can be used to either inform or be fed directly into subsequent analyses like visualization to depict clusters or networks. Our nonparametric approach, giving bias corrected mutual information (BCMI) estimates, has been automated along with associated statistical tests. BCMI is a very useful and general tool that can be widely used for exploratory data analysis.

Acknowledgment

This research has been supported by the Australian National Health and Medical Research Council grant 525453.

References

Cover, T.M. and Thomas, J.A. (2006) *Elements of Information Theory*. Wiley.

Dawy, Z., Goebel, B., Hagenauer, J., Andreoli, C., Meitinger, T. and Mueller, J.C. (2006) "Gene mapping and marker clustering using Shannon's mutual information," *EEE/ACM Transactions on Computational Biology and Bioinformatics* **3**: 47-56.

Efron, B. (2010) *Large-Scale Inference: Empirical Bayes Methods for Estimation* (Institute of Mathematical Statistics Monograph) Cambridge University Press.

Efron, B. and Gong, G. (1983) "A leisurely look at the bootstrap, the jackknife, and cross-validation," *American Statistician*, **37**: 36-48.

Pardy, C. (2013) *Mutual Information as an Exploratory Measure for Genomic Data with Discrete and Continuous Variables.* PhD thesis.

Reshef, D.N., Reshem Y.A., Finucane, H.K., Grossman, S.R., McVean, G., Turnbaugh, P.J., Lander, E.S., Mitzenmacher, M. and Sabeti, P.C. (2011) "Detecting novel associations in large data sets," *Science* **334**: 1518-1524.

Simon, N. and Tibshirani, R. (2011) "Comment on 'Detecting Novel Associations in Large Data Sets' by Reshef et. al: Science Dec 16, 2011.

Smoot, M., Ono, K., Ruscheinski, J., Wang, P.-L., Ideker, T. (2011) "Cytoscape 2.8: new features for data integration and network visualization**,"** *Bioinformatics*, **27**: 431–432.

Storey, J.D. and Tibshirani, R. (2003) "Statistical significance for genomewide studies." *Proceedings of the National Academy of Sciences* **100**(16): 9440-9445.

Wand, M.P. and Jones, M.C. (1995) *Kernel Smoothing*. Chapman & Hall.