5-11-2007

The Integrative Correlation Coefficient: a Measure of Cross-study Reproducibility for Gene Expressionea Array Data

Leslie M. Cope
Johns Hopkins University, cope@jhu.edu

Liz Garrett-Mayer
garrettm@musc.edu

Edward Gabrielson
Departments Oncology and Pathology, Johns Hopkins Medical Institute, egabriel@jhmi.edu

Giovanni Parmigiani
The Sydney Kimmel Comprehensive Cancer Center, Johns Hopkins University & Department of Biostatistics, Johns Hopkins Bloomberg School of Public Health, gp@jimmy.harvard.edu

Introduction

Whether the goal is to cross-validate research findings in independent data, to robustify results against technical differences between expression array platforms, or simply to increase sample size, multi-study analysis adds value to microarray experiments. However,

because of significant technical differences between microarray platforms, and because of differences in study design, it can be difficult to combine data. The ability to efficiently accumulate and integrate information from related genomic experiments will be critical to realizing the full benefit of the massive investment made on genomic studies.

We have developed a statistical measure of reproducibility that can be applied to individual genes, measured in two different studies. This statistic, which we call the Integrative Correlation Coefficient or Correlation of Correlations, borrows strength across many genes to estimate the strength of the relationship between expression values in the two studies. The integrative correlation was independently described by investigators at Stanford University, [5] and used to measure the cross-study reproducibility in microarrays, but not applied at the level of individual genes. The integrative correlation coefficient is self contained, depending only on expression values and not on supplemental sample data. It can be applied in situations where standard measures of correlation cannot, when the samples in the two studies are unrelated, and even when samples sizes differ. The integrative correlation coefficient has demonstrated its utility in several applications [6, 2] and herein we describe several important statistical properties as well.

Integrative correlation, and other tools for multi-study analysis are available in a software package called MergeMaid, [1] which is written in the R language [4] and included as part of the Bioconductor [3] bioinformatics software project.

Preliminaries: Definitions and Notation

Suppose that S_a and S_b are two microarray studies, with sample sizes of n_a and n_b respectively, and a total of m common genes. Assume (w.l.o.g. in the following) that all genes in each study are individually standardized to mean zero and variance one, and consider a particular gene x. The vector of expression values for gene x in study S_a is designated x_a, and A denotes the

$m-1 \times n_a$ data matrix for study S_a with gene x deleted. Notation for study S_b is, of course, identical.

Additionally, let $[c]_m$ denote the $m \times m$ matrix with every element equal to c, and let I_m denote the $m \times m$ identity matrix. Thus if v and w are two random vectors of length m then $[I_m - [1/m]_m]x$ returns $x - \overline{x}$, and $\text{cov}(x, y) = y^t[I_m - [1/m]_m]^2 x$.

Genes are already mean subtracted, but we will need to calculate several covariances over samples, so for convenience, let $\mathcal{I} = [I_m - [1/m]_m]^2$ and thus $\text{cov}(x, y) = y^t \mathcal{I} x$.

The Integrative Correlation Coefficient

It is not possible to correlate the expression levels of gene x in study S_a with expression levels in study S_b directly when the samples are not matched. However, the studies can be matched at the gene level, and individual samples, from the same or different studies, can be correlated over common genes. Thus our approach to measuring gene reproducibility is to construct virtual samples corresponding to each gene, in each study, and then correlate those across studies. The virtual samples are constructed very simply. Using study S_a to illustrate, the sample corresponding to gene x_a is a linear combination of the columns in the data matrices A, where the coefficients are just the expression intensities x_a and so can be written compactly in matrix form as Ax_a^t. Recollect that the data matrix A does not include gene x, and that x_a, as well as the rows of A are individually standardized in advance.

More formally, the integrative correlation coefficient for gene x in studies S_a and S_b is defined as

$$\frac{x_a A^t \mathcal{I} B x_b^t}{\sqrt{x_a A^t \mathcal{I} A x_a^t}\sqrt{x_b B^t \mathcal{I} B x_b^t}}.$$

In this formulation, it is easy to see that there are natural definitions for *integrative covariance* and *integrative variance* and as should be the case

$$\text{i}cor(x_a, x_b) = \frac{\text{i}cov(x_a, x_b)}{\sqrt{\text{i}var(x_a)}\sqrt{\text{i}var(x_b)}}.$$

The integrative covariance of gene x is thus

$$x_a A^t \mathcal{I} B x_b^t,$$

and the integrative variance of x_a is

$$x_a A^t \mathcal{I} A x_a^t.$$

The central object in the integrative variance of x_a is the $n_a \times n_a$ sample covariance matrix $A^t \mathcal{I} A$. In the covariance expression, there is an analagous $n_a \times n_b$ cross-covariance matrix $A^t \mathcal{I} B$. The i, j-th entry of the cross-covariance matrix is the covariance between sample i of study a and sample j of study b.

To interpret the integrative correlation, we first note the key assumption: similar samples should have well correlated expression profiles even across studies. This should be true when samples are correlated across the full set of common genes, and for the reproducible set, it should be true at the individual gene level as well. Thus we assume that a well measured gene will weigh like samples We will also assume that the studies are selected to include comparable cohorts. Thus, if gene x is reproducibly measured in the two studies, then the linear combinations that x defines on the samples in each study should weigh like subclasses of the cohorts in the same way. We do not believe that the assumptions implicit in this argument need to be explicitly verified before applying an integrative correlation analysis. If there is a *Fundamental Theory of Expression Array Analysis* it is that phenotype is reflected in the expression profile for a sample. If this is not true, then the failure of the integrative correlation is the least of our problems. Nonetheless, later on in this paper we do consider ways of determining when two studies are appropriate for integrative correlation analysis.

This approach to cross-study reproducibility has clear strengths. First of all of course, it permits the use of a standard correlation coeffient in a setting where it would not otherwise be applicable. The integrative correlation coefficient is closely related to classical canonical correlation analysis as well. We will exploit these connections to develop statistical theory for the measure. The integrative correlation does not incorporate phenotypic information directly, but only insofar as phenotype is reflected in the expression profile. Thus genes can be filtered on the basis of integrative correlation without biasing the selection process toward subsequent, higher level analyses. And although it is computationally expensive to calculate the integrative correlation exactly for a large number of genes, an accurate and efficient approximation is available.

Some Statistical Properties of the Integrative Correlation Coefficient

In this section, we describe theoretical properties of the integrative correlation coefficient. Some of the most straightforward results are mentioned briefly for completeness, and are not proved here. Along the way we point out where open questions remain.

The integrative correlation is invariant to re-ordering of data columns, or data rows (as long as rows are re-ordered in tandem in both studies).

After sample covariance values are calculated, the vectors x_a and x_b can be re-scaled without changing integrative correlation. In particular, each can be scaled so that the integrative variances are equal to 1, in which case the integrative correlation takes the simple form $x_a A^t \mathcal{I} B x_b^t$.

The largest possible integrative correlation coefficient can be calculated as the largest canonical correlation coefficient of the two datsets, standardized by gene. Likewise multiple regression can be used to calculate the largest integrative correlation coefficient that can obtained by changing a single value.

The Null Distribution

The idea behind our null distribution is to keep all genes and samples as they are, and add new *null* genes one at a time, calculating the integrative correlation against the background of the actual data. The null genes are to have no correlation across studies, and so can be generated in each study as independent random variates.

The matrix formulation of IntCor facilitates efficient sampling from the null distribution. The within-study and cross=study empirical sample covariance matrices are used as is, while independent standard normal vectors v_a and v_b are substituted for x_a and x_b. The computational cost of each iteration is proportional to the squared sample size rather than the square of the number of genes.

Experiments with uniform, exponential and normal distribution random variables for this, and all give *exactly* the same distribution of integrative correlations. The qqplots are straight lines.

Conditioning on samples, the null distribution of CorCor scores has mean

0 and variance

$$\sigma^2 = \frac{\sum_{ij} \rho_{ij}^2}{n_a \times n_b} - \mu(\rho_{ij})$$

where the ρ_{ij} are the expected cross-study sample correlations.If sample size $n_a \times n_b \to \infty$ then the asymptotic null distribution is Normal.

Outstanding problems and ongoing work

There are several outstanding problems associated with the integrative correlation.

- What is the most natural and useful generalization to more than two studies? We have considered averaging all pairwise correlations of correlations...
- Is there a natural inter-study correlation coefficient, describing in a single number the amount of information that is shared by two independent studies measured on the same variables. The largest possible integrative correlation may be such a coefficient.

Extensions

There are several places to go with this.

Calculation

The matrix formulation may significantly reduce computational costs. Sample covariance and cross-covariance matrices are calculated on the deleted data sets. However if there are a lot of genes we can expect that these matrices will not be sensitive to the inclusion of one gene more or less. In many circumstances it is probably fine to use the entire, undeleted dataset as the sole, common cross-covariance matrix. There is still a lot of computation, but it may be possible to find other good approximations that further reduce the cost.

If the matrices are the same for all genes, then it is easy to calculate all integrative correlations at once. The vector of coefficients is obtained as

$$\mathrm{d}iag(\sqrt{\Sigma_a^{-1}} A A^t \mathcal{I} B B^t \sqrt{\Sigma_b^{-1}}),$$

where $\Sigma_a = diag(AA^t\mathcal{I}AA^t)$, and since Σ is diagonal and positive, the square root simply means the element-wise square root.

Generalizations

Think of the sample covariance and sample cross-covariance matrices as defining intersample similarity. Substitute other measures of similarity, like sample correlation and sample cross-correlation. Or use some phenodata to define new within-study and cross-study sample similarity measures and substitute those.

Put matrix and linear combination versions together to think about how this works. A gene defines a linear combination on samples, or equivalently, a kind of bi-quadratic form on sample covariances, as described in the matrix version. Ideally, an alternative similarity measure, would correspond to a data matrix with "alternative" gene expression values, where the new sample similarities describe the sample information shared in the new expression values. We **should** use the alternative expressions to define the linear combination in this setting, but since they don't really exist, we have to work with what we do have. The genes with high integrative correlation would then be those whose (real) expressions happen to contain the information captured by the virtual expressions.

Internal Integrative Correlation

Quishan Tao applied integrative correlation to a single study, comparing two genes to one another, but borrowing strength across the entire set. The formulation is just the same as above, except that in this instance there is only one data matrix, so $A = B$, and the similarity matrix is the same for both the integrative covariance and the integrative variances.

It is likelythe internal integrative correlation can be exploited. The key is that genes with high internal integrative correlation share information about the subjects profiled in the study; information that is sufficiently global as to show up clearly in the *other* genes as well.

References

[1] Leslie Cope, Xiaogang Zhong, Elizabeth Garrett, and Giovanni Parmigiani. Mergemaid: R tools for merging and cross-study validation of gene

expression data. *Stat Appl Genet Mol Biol*, 3(1):Article29, 2004.

[2] Leslie Cope, Xiaogang Zhong, Elizabeth Garrett-Mayer, Edward Gabrielson, and Giovanni Parmigiani. Cross-study validation of the molecular profile of brca1-linked breast cancers. *Unpublished Manuscript*, 2004.

[3] Robert Gentleman. BioConductor: open source software for bioinformatics. `http://www.bioconductor.org`, 2003.

[4] Ross Ihaka and Robert Gentleman. R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3):299–314, 1996.

[5] Jae K Lee, Kimberly J Bussey, Fuad G Gwadry, William Reinhold, Gregory Riddick, Sandra L Pelletier, Satoshi Nishizuka, Gergely Szakacs, Jean-Phillipe Annereau, Uma Shankavaram, Samir Lababidi, Lawrence H Smith, Michael M Gottesman, and John N Weinstein. Comparing cdna and oligonucleotide array data: concordance of gene expression across platforms for the nci-60 cancer cells. *Genome Biology*, 4:doi:10.1186/gb–2003–4–12–r82, 2003.

[6] Giovanni Parmigiani, Elizabeth S. Garrett-Mayer, Ramaswami Anbazhagan, and Edward Gabrielson. Cross-study comparison of gene expression data sets for the molecular classification of lung cancer. *Clinical Cancer Research*, 10(9):in press, 2004.