

Gene expression

Eigen- R^2 for dissecting variation in high-dimensional studies

Lin S. Chen¹ and John D. Storey^{1,2,*}

¹Lewis-Sigler Institute and ²Department of Molecular Biology, Princeton University, Princeton, NJ 08544, USA

Received on March 18, 2008; revised on June 3, 2008; accepted on August 1, 2008

Advance Access publication August 20, 2008

Associate Editor: Joaquin Dopazo

ABSTRACT

Summary: We provide a new statistical algorithm and software package called ‘eigen- R^2 ’ for dissecting the variation of a high-dimensional biological dataset with respect to other measured variables of interest. We apply eigen- R^2 to two real-life examples and compare it with simply averaging R^2 over many features.

Availability: An R-package `eigenR2` is available at <http://www.genomine.org/eigenR2/> and will be made publicly available via Bioconductor.

Contact: jstorey@princeton.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 INTRODUCTION

A common goal in the analysis of high-dimensional biological studies is to decompose the variation of thousands of measured features in terms of some other variables. In statistical parlance, the measured features are considered a set of related ‘response variables’ and the other variables used to explain their variation are called ‘independent variables’.

A common application of this goal is in dissecting the variation of transcriptional levels of thousands of genes in terms of relevant biological variables. For example, Jin *et al.* (2001) estimated the contributions of sex, genotype and age to transcriptional variation in *Drosophila melanogaster*. They found that expression variation is mostly explained by sex, genotype and their interactions, and less explained by age. Ross *et al.* (2000) characterized the transcriptional variation of thousands of genes across human cancer cell lines. Brem *et al.* (2002) dissected the variation of expression in yeast according to genotype, based on recombinant lines derived from two distinct isogenic strains. Similarly, Morley *et al.* (2004) analyzed genome-wide variation in human gene expression in terms of underlying genetic variation. The genome-wide transcription variation explained by population structure has been estimated in the teleost fish (Oleksiak *et al.*, 2002) and humans (Storey *et al.*, 2007).

In all of these studies, one can think of each feature as being a response variable, where a key summary statistic is the proportion of variation among the thousands of response variables explained by the independent variables of interest. For a single response variable, the proportion of variation explained by independent variables is usually accomplished by calculating R^2 . This is computed as the ratio of the variance of the fitted model to the variance of

the response variable. For example, *heritability* is the R^2 value according to a particular model of the quantitative trait based on genotypes.

With thousands of response variables, one can calculate R^2 values for each one, resulting in thousands of these values. Even though it is reasonable to simply plot the distribution of these R^2 values, sometimes it is also desirable to calculate an average R^2 , so that statements can be made about the overall proportion of variation in the response variables explained by variables of interest. One obvious choice is to simply take the mean of the R^2 values, which we refer to as mean- R^2 . However, this measure may be vulnerable to technical artifacts in the data. For example, in a gene expression study, many genes will not be expressed. For such genes, the expression measurements are low, resulting in the ratio of the mean expression to the SD being unstable. In this case, the R^2 values of these genes would fluctuate wildly. This has been recognized as an issue in testing genes for differential expression, and usually a small constant is added to the estimated variance of the genes to down weight the genes that are likely not expressed (Tusher *et al.*, 2001). Another solution that has been implemented is to eliminate genes on the arrays before any analysis is done by a presence/absence call.

Here, we introduce a new statistical algorithm and software package for calculating a quantity called eigen- R^2 . The eigen- R^2 quantity is computationally efficient to calculate, and it is robust to the issues described above. We show, for example, that the eigen- R^2 values in an expression study are robust to different levels of gene-filtering based on a presence/absence call, whereas mean- R^2 is strongly affected by gene-filtering.

2 METHODS

Suppose that the high-dimensional biological dataset is organized as an $m \times n$ matrix \mathbf{Y} , where the rows of \mathbf{Y} represent different response variables and the columns represent different observations. For example, in a gene expression study, the rows of \mathbf{Y} are the m genes and the columns of \mathbf{Y} are the n arrays. Suppose also that an additional variable $\mathbf{z} = (z_1, z_2, \dots, z_n)^T$ has been measured. For example, \mathbf{z} could be a clinical variable, genotype or treatment. If $\mathbf{y}_i = (y_{i1}, y_{i2}, \dots, y_{in})^T$ is the data for response variable i , then we may fit a model of \mathbf{y}_i on \mathbf{z} to obtain fitted values $\hat{\mathbf{y}}_i$. The proportion of variation in \mathbf{y}_i that is explained by \mathbf{z} is then:

$$R_{\mathbf{y}_i}^2 = \frac{\hat{\sigma}_{\hat{\mathbf{y}}_i}^2}{\hat{\sigma}_{\mathbf{y}_i}^2} = \frac{\sum_{j=1}^n (\hat{y}_{ij} - \bar{y}_i)^2}{\sum_{j=1}^n (y_{ij} - \bar{y}_i)^2},$$

where \bar{y}_i is the mean of $\hat{\mathbf{y}}_i$ and \bar{y}_i is the mean of \mathbf{y}_i . It then follows that mean- R^2 is the average of these across all response variables, $\sum_{i=1}^m R_{\mathbf{y}_i}^2 / m$.

Instead of taking the average of the $R_{\mathbf{y}_i}^2$, we have developed an approach to employ principal components analysis to measure an overall R^2 . In principal

*To whom correspondence should be addressed.

components analysis, a singular value decomposition is applied to the data matrix, decomposing \mathbf{Y} into the following: $\mathbf{Y} = \mathbf{UDV}^T$, where the matrices \mathbf{U} and \mathbf{V} are column orthogonal so that $\mathbf{U}^T\mathbf{U} = \mathbf{V}^T\mathbf{V} = \mathbf{I}$ and \mathbf{D} is a diagonal matrix. The columns of \mathbf{V} are the right eigenvectors and the columns of \mathbf{U} are the left eigenvectors. We are particularly interested in the right eigenvectors, because these represent aggregated trends in the response variables. Specifically, the first column of \mathbf{V} is the linear combination of all response variables that explains the most variation in the data, called the first right eigenvector. The second column of \mathbf{V} is the linear combination of all response variables that explains the most variation in the data once the first eigenvector has been removed, and so on. The proportion of total variation captured by the i -th eigenvector is $\pi_i = d_i^2 / \sum_{l=1}^L d_l^2$, where d_i is the eigenvalue of the i -th eigenvector, which is obtained from the i -th diagonal entry of \mathbf{D} .

When \mathbf{Y} is a matrix of expression data, the right eigenvectors have been called ‘eigen-genes’. Alter *et al.* (2000) first proposed this terminology and showed that principal components analysis is a useful tool for identifying major trends in expression data. The top few eigen-genes tend to capture biologically relevant trends present in a number of genes. In a more general setting, one could call the columns of \mathbf{V} ‘eigen-response-variables’.

Let \mathbf{v}_i be the i -th column of \mathbf{V} and let $\hat{\mathbf{v}}_i$ be the fitted values when modeling \mathbf{v}_i in terms of \mathbf{z} . For each of these, we can calculate an R^2 value, denoted by $R_{\hat{\mathbf{v}}_i}^2$. Since π_i of the total variation in the data is explained by \mathbf{v}_i , $R_{\hat{\mathbf{v}}_i}^2$ should be weighted by π_i . Additionally, since each pair of eigen-response-variables is uncorrelated, the variation explained by \mathbf{z} in \mathbf{v}_i is orthogonal to the variation explained by \mathbf{z} in \mathbf{v}_j where $j \neq i$. Therefore, as an overall measure of R^2 , we proposed to take the average of the $R_{\hat{\mathbf{v}}_i}^2$, weighted by their respective π_i :

$$\text{eigen-}R^2 = \sum_{i=1}^n \pi_i R_{\hat{\mathbf{v}}_i}^2.$$

The following algorithm summarizes the eigen- R^2 calculation:

Step 1. Let \mathbf{Y} be an $m \times n$ matrix, where the rows of \mathbf{Y} represent different response variables and have been mean centered, and the columns represent different observations. Use singular value decomposition to decompose the matrix of response variables as $\mathbf{Y} = \mathbf{UDV}^T$.

Step 2. For each column of \mathbf{V} , denoted by \mathbf{v}_i , fit the user-specified model of \mathbf{v}_i on the independent variable(s) \mathbf{z} to obtain fitted values $\hat{\mathbf{v}}_i$, $i = 1, 2, \dots, n$. Calculate the R^2 value of this model fit as described above to obtain $R_{\hat{\mathbf{v}}_i}^2$.

Step 3. Calculate the proportion of variation explained by \mathbf{v}_i with $\pi_i = d_i^2 / \sum_{l=1}^L d_l^2$.

Step 4. Calculate the overall R^2 value as $\text{eigen-}R^2 = \sum_{i=1}^n \pi_i R_{\hat{\mathbf{v}}_i}^2$.

Straightforward linear algebra shows that using models that can be fitted by applying a linear operator (e.g. ordinary least squares regression, least squares regression with spline functions), eigen- R^2 can equivalently be written as follows:

$$\text{eigen-}R^2 = \sum_{j=1}^m w_j R_{\hat{\mathbf{v}}_j}^2, \text{ where } w_j = \frac{\hat{\sigma}_{y_j}^2}{\sum_{k=1}^m \hat{\sigma}_{y_k}^2}.$$

In this case, it follows that eigen- R^2 can be written as a weighted average of the feature-specific $R_{\hat{\mathbf{v}}_j}^2$ values, where the weights correspond to the response variable’s contribution to the overall distribution of baseline variances. It also follows from this equivalence that if the rows of \mathbf{Y} are not only mean-centered, but also scaled to have unit SD, then eigen- R^2 is equal to mean- R^2 . In the Supplementary Material, we prove the above equivalence and compare the eigen- R^2 measure to two other methods for determining the weights.

3 SOFTWARE

We have developed an R library called `eigenR2` to perform this algorithm. This R library includes a plot function to graphically display the information on the eigen-response-variables. By default,

Table 1. Eigen- R^2 and mean- R^2 values for the proportion of expression variation attributed to within- and between-populations, from Storey *et al.* (2007)

No. of probe sets	8746	5969	5194	4646	$\frac{\text{Max}-\text{Min}}{\text{Min}}$ (%)
Individual eigen- R^2	0.399	0.414	0.418	0.411	4.6
Individual mean- R^2	0.197	0.248	0.264	0.274	39.0
Population eigen- R^2	0.071	0.075	0.076	0.076	7.0
Population mean- R^2	0.045	0.051	0.054	0.055	22.1

Each column represents a different threshold for filtering probe sets during preprocessing. Technical variation was not removed before calculating R^2 , so this variable explains the remaining variation in the data.

the R package estimates eigen- R^2 with models fitted by least squares, while it also has the flexibility that allow users to define their own R^2 estimating functions to obtain eigen- R^2 estimates.

The software offers an option to adjust the R^2 calculations to accommodate small sample sizes (small n) by replacing R^2 with $1 - (1 - R^2) \times (n - d_0) / (n - d)$, where d_0 and d are the degrees of freedom spent in fitting the null model and the target model, respectively. By default, the null model is a model with only intercept and $d_0 = 1$. The adjusted estimator has been shown to be an unbiased estimator of the true value (Schroeder *et al.*, 1986). The software also provides the option to utilize only those eigenvectors that explain more variation than would be expected by chance. When the significance threshold is specified, the function estimates the P -value for each eigen-gene (Leek and Storey, 2007), and only selects the ones pass the threshold to estimate eigen- R^2 . This acts as an additional mechanism for de-noising the overall eigen- R^2 estimate.

4 EXAMPLES

4.1 Expression variation within and among populations

Storey *et al.* (2007) recently analyzed patterns of natural gene expression variation in B lymphoblastoid cells derived from 16 HapMap individuals of European and African ancestry, eight from each respective population. They characterized the gene expression variation to within and among these populations for each gene by utilizing R^2 calculations, and also tested for genes differentially expressed within and among populations. We employed the eigen- R^2 method and software on this dataset to produce a genome-wide summary of the within-population and inter-population contributions to genome-wide expression variation.

In Table 1, we present a comparison of the eigen- R^2 and mean- R^2 values over varying numbers of probe sets obtained from different filtering thresholds available in the `afey` package for calling a probe set ‘expressed’. The thresholds were set according to probe sets being expressed in at least 0%, 25%, 50% and 75% of the samples, respectively. The results are shown in Table 1. Because we did not remove the affect of technical variation from these quantities, it is possible as a proof of concept to compare the performance of both estimators according to different levels of technical noise. Over all filtering thresholds, the eigen- R^2 quantities are higher than mean- R^2 for both within- and between-population variation, which implies that the mean- R^2 are reduced more by technical variation. Also, the eigen- R^2 quantities are relatively stabler than mean- R^2 quantities across various filtering thresholds.

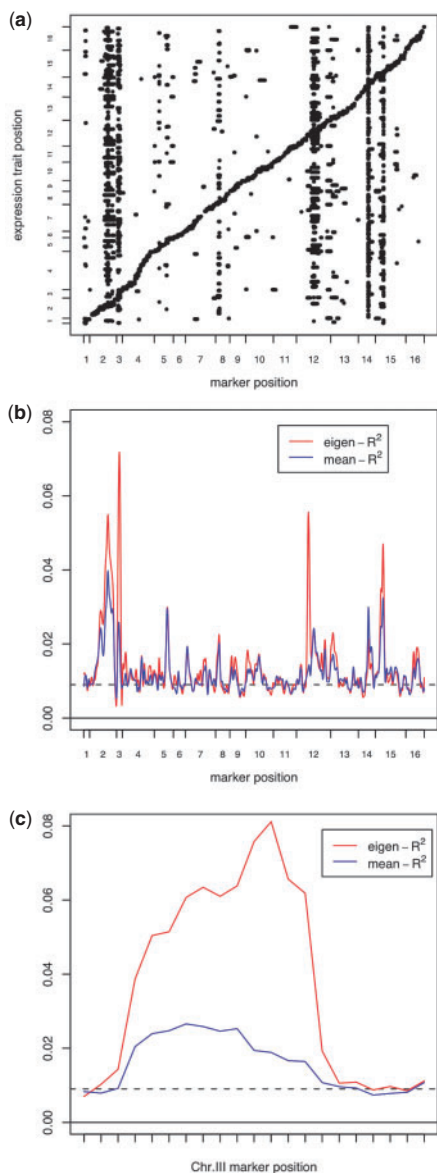


Fig. 1. (a) Tests for linkage among all gene expression trait and marker pairs. A dot indicates significant linkage. (b) Genome-wide comparison of eigen- R^2 and mean- R^2 values. The dashed horizontal line shows the expected R^2 value under no linkage signal. (c) Eigen- R^2 and mean- R^2 values on Chromosome III, which contains the important *MAT* locus.

4.2 Genetic dissection of transcriptional variation

Brem *et al.* (2002) explored the genetic architecture of budding yeast by analyzing data from an experimental cross between two strains of *Saccharomyces cerevisiae*. The parental strains are haploid derivatives of a standard laboratory strain (BY) and a wild isolate from a California vineyard (RM). For the 112 segregants, expression levels of ~ 6000 genes were measured and genetic markers were identified with oligonucleotide microarrays. We performed linkage analysis of the expression levels similarly to Brem *et al.* (2002). At a P -value cutoff of 5×10^{-7} , about 9000 out of the ~ 19 million of gene-marker combinations show significant linkage. Figure 1a plots the significantly linked gene expression trait positions against the linked marker positions on the yeast genome. We estimated the eigen- R^2 and mean- R^2 values at each locus, shown in Figure 1b. It can be seen from these plots that both quantities capture the linkage hotspots well. However, eigen- R^2 tends to capture more signal, particularly on Chromosomes II, III, XII, XIII and XIV. The *MAT* mating locus on Chromosome III, which has been shown to play an important role in gene expression (Brem *et al.*, 2002), has the highest eigen- R^2 of 8.1%, while the mean- R^2 at that locus is 1.9% (Fig. 1c).

Conflict of Interest: none declared.

REFERENCES

- Alter, O. *et al.* (2000) Singular value decomposition for genome-wide expression data processing and modeling. *Proc. Natl Acad. Sci. USA*, **97**, 10101–10105.
- Brem, R.B. *et al.* (2002) Genetic dissection of transcriptional regulation in budding yeast. *Science*, **296**, 752–755.
- Jin, W. *et al.* (2001) The contributions of sex, genotype and age to transcriptional variance in *Drosophila melanogaster*. *Nat. Genet.*, **29**, 389–395.
- Leek, J.T. and Storey, J. (2007) Capturing heterogeneity in gene expression studies by “surrogate variable analysis”. *PLoS Genet.*, **3**, e161.
- Morley, M. *et al.* (2004) Genetic analysis of genome-wide variation in human gene expression. *Nature*, **430**, 743–747.
- Oleksiak, M.F. *et al.* (2002) Variation in gene expression within and among natural populations. *Nat. Genet.*, **32**, 261–266.
- Ross, D.T. *et al.* (2000) Systematic variation in gene expression patterns in human cancer cell lines. *Nat. Genet.*, **24**, 208–209.
- Schroeder, L.D. *et al.* (1986) *Understanding Regression Analysis: An Introductory Guide*. Sage Publications, Thousand Oaks, CA. *Quantitative Applications in the Social Sciences*.
- Storey, J.D. *et al.* (2007) Gene expression variation within and among human populations. *Am. J. Hum. Genet.*, **80**, 502–509.
- Tusher, V.G. *et al.* (2001) Significance analysis of microarrays applied to the ionizing radiation response. *Proc. Natl Acad. Sci. USA*, **98**, 5116–5121.