12-30-2015

EntropyExplorer: An R Package for Computing and Comparing Differential Shannon Entropy

Kai Wang  
*Georgia Southern University, kwang@georgiasouthern.edu*

Charles A. Philips  
*University of Tennessee, Knoxville*

Arnold M. Saxton  
*University of Tennessee Institute of Agriculture, Knoxville*

Michael A. Langston  
*University of Tennessee, Knoxville*

Follow this and additional works at: [https://digitalcommons.georgiasouthern.edu/compsci-facpubs](https://digitalcommons.georgiasouthern.edu/compsci-facpubs)  
Part of the [Computer Sciences Commons](https://digitalcommons.georgiasouthern.edu/compsci-facpubs)

Recommended Citation  
[https://digitalcommons.georgiasouthern.edu/compsci-facpubs/53](https://digitalcommons.georgiasouthern.edu/compsci-facpubs/53)
EntropyExplorer: an R package for computing and comparing differential Shannon entropy, differential coefficient of variation and differential expression

Kai Wang¹, Charles A. Phillips¹*, Arnold M. Saxton² and Michael A. Langston¹

Abstract

Background: Differential Shannon entropy (DSE) and differential coefficient of variation (DCV) are effective metrics for the study of gene expression data. They can serve to augment differential expression (DE), and be applied in numerous settings whenever one seeks to measure differences in variability rather than mere differences in magnitude. A general purpose, easily accessible tool for DSE and DCV would help make these two metrics available to data scientists. Automated p value computations would additionally be useful, and are often easier to interpret than raw test statistic values alone.

Results: EntropyExplorer is an R package for calculating DSE, DCV and DE. It also computes corresponding p values for each metric. All features are available through a single R function call. Based on extensive investigations in the literature, the Fligner-Killeen test was chosen to compute DCV p values. No standard method was found to be appropriate for DSE, and so permutation testing is used to calculate DSE p values.

Conclusions: EntropyExplorer provides a convenient resource for calculating DSE, DCV, DE and associated p values. The package, along with its source code and reference manual, are freely available from the CRAN public repository at http://cran.r-project.org/web/packages/EntropyExplorer/index.html.

Keywords: Differential Shannon entropy, Differential coefficient of variation, Differential expression, Statistical tests

Background

Shannon entropy (SE) and coefficient of variation (CV) are used to measure the variability or dispersion of numerical data. Such variability has potential utility in numerous application domains, perhaps most notably in the analysis of high throughput biological data. Variability has been applied, for example, to study gene expression data in the context of human disease [1]. Increased entropy in particular, in both gene expression and protein interaction data, has been observed to be a characteristic of cancer [2]. Numerous other examples typify the utility of entropy [3–8] and coefficient of variation [9–12].

Shannon entropy is famously rooted in information theory [13]. To avoid confusion, we emphasize that we use the term “differential entropy” to denote a difference between two Shannon entropy values. This is distinct from information-theoretic terminology, in which “differential entropy” often means the entropy of a continuous, rather than a discrete, random variable [14].

We are particularly interested in differential analysis. In [15], we studied differential Shannon entropy (DSE) and differential coefficient of variation (DCV), and found them highly effective in identifying genes of potential interest not found by differential expression (DE) alone. DSE and DCV are applicable to other types of biological data as well, such as that produced by RNA-Seq technologies, although the usual caveats about careful interpretation apply. The usefulness of DSE and DCV is of course

*Correspondence: cphill25@tennessee.edu
1 Department of Electrical Engineering and Computer Science, University of Tennessee, Knoxville, TN 37996-2250, USA
Full list of author information is available at the end of the article

© 2015 Wang et al. This article is distributed under the terms of the Creative Commons Attribution 4.0 International License (http://creativecommons.org/licenses/by/4.0/), which permits unrestricted use, distribution, and reproduction in any medium, provided you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license, and indicate if changes were made. The Creative Commons Public Domain Dedication waiver (http://creativecommons.org/publicdomain/zero/1.0/) applies to the data made available in this article, unless otherwise stated.
not limited to biological data. They may be applied to any numerical data for which normalized measures of differential variability are relevant.

**Implementation**

*EntropyExplorer* is implemented in R [16]. All features are wrapped into a single function call, which takes as input up to eight arguments. Two of these arguments are numerical matrices, with identical labels for each row. The output is a matrix with two, three or five columns that contains in each row two SE, CV or mean values; a DSE, DCV or DE value; and/or two p values, one raw and one adjusted. Output rows can be sorted by value, raw p value or adjusted p value, and can be filtered to show only the top-ranked rows.

Permutation testing for DSE is accomplished with the help of the R function *sample.int*. The default number of tests to be employed is set to 1000, which the user can override. The p value for DCV is calculated by applying the Fligner-Killeen test for homogeneity of variances, implemented via the R function *fligner.test*, to the log-transform of the input data. The R function *t.test* is used to find a p value for DE. Adjusted p values are calculated using the *p.adjust* function in R, which provides false discovery rate and multiple testing corrections. A more thorough explanation of p value calculations is provided in the discussion section.

*EntropyExplorer* checks that all matrix entries are positive. This is because calculations of a DSE value/p value and a DCV p value involve taking logarithms, which are undefined on data containing zeros or negative values. Also, CV becomes less meaningful when means approach zero or are negative. Experimental data may be noisy, however, and so *EntropyExplorer* provides mechanisms to handle non-positive values. An optional two-value argument permits the user to add a positive bias to all elements of one or both matrices prior to performing any other calculations. The argument can also be set to make this adjustment automatically, based on the least non-positive value in each matrix.

**Metrics**

Let \( x_1, x_2, \ldots, x_n \) represent a list of \( n \) positive numbers, and let \( x = \sum_{i=1}^{n} x_i \) denote their sum. The Shannon entropy of this list is

\[
SE = \frac{-\sum_{i=1}^{n} \frac{x_i}{x} \log_2 \frac{x_i}{x}}{\log_2 n}.
\]

The coefficient of variation is

\[
CV = \frac{s}{|\bar{x}|}
\]

where \( \bar{x} = x/n \) is the sample mean and \( s = \sqrt{\frac{\sum_{i=1}^{n} (x_i-\bar{x})^2}{n-1}} \) is the sample standard deviation. Given two such lists of positive numbers with Shannon entropies \( SE_1 \) and \( SE_2 \), coefficients of variation \( CV_1 \) and \( CV_2 \), and means \( \bar{x}_1 \) and \( \bar{x}_2 \), \( DSE = |SE_1 - SE_2| \), \( DCV = |CV_1 - CV_2| \), and \( DE = |\bar{x}_1 - \bar{x}_2| \).

Shannon entropy falls in the range \([0, 1]\); DSE therefore also falls in the range \([0, 1]\). Lower (higher) SE corresponds to more (less) variability. CV falls in the range \([0, \infty)\); DCV therefore also has a range of \([0, \infty)\).

**Application**

*EntropyExplorer* is invoked as follows:

> *EntropyExplorer*(m1,m2,dmetric="dse",otype="v",ntop=10)

SE(expm1) SE(expm2) SE(expm1)-SE(expm2)
228245_s_at 0.4585632 0.9749417 -0.5163785
205220_at 0.4347483 0.9494563 -0.5147080
213711_at 0.4658234 0.9705819 -0.5047585
209242_at 0.4757616 0.9597454 -0.4839837
203908_at 0.4971761 0.9460891 -0.4489130
205030_at 0.4183439 0.8599655 -0.4416216
227282_at 0.5773223 0.9901547 -0.4128224
205067_at 0.5584833 0.9406934 -0.3822100
223623_at 0.6033164 0.9801690 -0.3768526
203824_at 0.5562821 0.9249965 -0.3687144

**Fig. 1** The output of *EntropyExplorer* on breast cancer data. The numerical matrices m1 and m2 have been read into R. The function call has specified “dse” for differential Shannon entropy, “v” for value, and 10 to return the top 10 values.
two metrics. For DCV p values, we observe that 11 tests of equal relative variation were compared in [21], with the conclusion that the Fligner-Killeen test [22] is usually the most appropriate. It strikes a balance between type I and type II errors, and is robust to non-normal distributions.

Obtaining reliable p values for DSE proved much more challenging. We found no known method in the literature specific to DSE p values. We therefore investigated the extent to which SE is correlated to variance. A high correlation would suggest that they may be proxies for each other, in which case the p value of an F-test or some derivation thereof might serve as suitable estimate of the DSE p value. Unfortunately, correlations between SE and variance, or between SE and a function of variance, were not high enough to justify using one as a surrogate for the other. Table 1 shows the correlation between SE and variance $V$, and between SE and the function $\frac{1}{2} \ln (2\pi e V)$ as an attempt to linearize the relationship, using the 16 datasets from [15]. The only notably high correlation is found in the obesity dataset. The obesity data, however, contains a large number of missing values, rendering the high correlation less reliable. We conclude that standard statistical tests related to variance do not appear suitable for testing DSE.

We also examined the distribution of DSE on the 16 datasets, with the goal of empirically determining a suitable reference distribution for DSE. From this, we could then estimate p values analytically. We applied the Kolmogorov–Smirnov (KS) test to compare the DSE distribution of each dataset to some of the more common reference distributions, such as normal, F, t, and Chi square. When performing a KS test, p values can be overly sensitive to deviations from the reference distribution [23], so a D-statistic value below 0.1 was used to identify matching distributions. In our experiments, only the Parkinson's dataset produced a D-statistic below 0.1 when tested against a normal or standardized t distribution (Table 2). Figure 3 shows a sample distribution of DSE, in this case using prostate cancer data.
We conclude from this that none of the distributions tested are close enough approximations to the observed DSE distribution to be used as a proxy for obtaining p values. Thus, without a known distribution function or suitable surrogate, we resort to resampling in order to obtain reliable DSE p values. While computationally demanding, the following permutation test makes no assumptions about the underlying distribution of the data. Given two lists of numbers, containing \( n_1 \) and \( n_2 \) numerical elements respectively, we first calculate their DSE and then create a new list \( A \) containing all \( n_1 + n_2 \) numbers from the two lists. Next we randomly permute the elements of \( A \), then recalculate DSE, treating the first \( n_1 \) elements of \( A \) as one list and the last \( n_2 \) elements of \( A \) as a second list. The resultant p value is simply the proportion of all recalculated DSEs that are at least as extreme as the original DSE.

In addition to raw p values, *EntropyExplorer* also calculates p values adjusted for multiple testing. A user can choose to adjust based on FDR, Holm or another multiple-testing adjustment.

**Conclusions**
We have produced *EntropyExplorer*, an R package for calculating differential Shannon entropy, differential coefficient of variation and differential expression. This package also calculates raw and adjusted p values for each metric. These measures have been shown to complement one another [15], making this package an effective tool for users in search of more expansive suites of differential analysis methods.

**Availability and requirements**

**Project name:** *EntropyExplorer*.  
**Project home page:** [http://cran.r-project.org/web/packages/EntropyExplorer/index.html](http://cran.r-project.org/web/packages/EntropyExplorer/index.html).  
**Operating system(s):** Platform independent.  
**Programming language:** R.  
**Other requirements:** R version 3.0 or later is recommended.
References