MEASURING SURPRISE WITH RELATIVE ENTROPY

Sevda GÜRSAKAL*, Dilek MURAT ** Necmi GÜRSAKAL ***

* Assist. Prof. Dr., Uludağ University, Faculty of Economics and Administrative Science, Department of Econometrics E-Mail: sdalgic@uludag.edu.tr
** Dr., Uludağ University, Faculty of Economics and Administrative Science, Department of Econometrics
*** Prof. Dr., Uludağ University, Faculty of Economics and Administrative Science, Department of Econometrics

Copyright © 2015 Sevda GÜRSAKAL, Dilek MURAT and Necmi GÜRSAKAL. This is an open access article distributed under the Eurasian Academy of Sciences License, which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

ABSTRACT
In recent times entropy concept has been used in many fields such as neural networks, pattern recognition, and statistics for the goodness of fit tests. Surprise attracts human attention and from marketing and advertisement to education attracting human attention is very important. In a Bayesian context, this paper tries to explain how to measure “surprise” by using relative entropy concept and discusses the various implications of this measurement.

Keywords: Relative entropy, Kullback-Leibler divergence, Bayesian statistics, R.

1. INTRODUCTION
“Entropy” is a measure of disorder, complexity and randomness. This concept has been used in physics, chemistry, information theory and statistics. Entropy is not easy to grasp and have a large number of interpretations, explications, applications (Benguigui, 2013: 304). For example, complexity of graphs can be measured by entropy concept (Mowshowitz and Dehmer, 2012: 562). Dictionaries have many definitions of entropy: For a closed system a quantitative measure of the amount of thermal energy, a measure of the disorder or randomness in a closed system, a measure of the loss of information for a transmitted message are all definitions of entropy.

Engineers in the late-18th and the 19th centuries were interested in heat engines and for them heat was a kind of weightless fluid—called caloric—that flowed from a hot body to a cold body (Ben-Naim, 2010: 1). Rudolf Clausius (1822–1888), in the context of thermodynamics is the father of entropy and published a series of papers on the theory of heat and its
applications to engines (Benguigui, 2013: 305). Clausius had defined the entropy quantity in terms of heat and temperature.

According to Clausius, there can be no process the sole result of which is a flow of energy from a colder to a hotter body. Also Nicolas Leonard Sadi Carnot (1796–1832) tried to answer the question of how much work can be generated from a given amount of heat. The Second Law of the Thermodynamics has many formulations but mainly it states that, “in a closed system the entropy cannot decrease, it remains constant or increases”. For Clausius, in any spontaneous process, occurring in an isolated system, the entropy never decreases”. Water molecules in an ice cube are ordered. But when they are heated up and transformed into gas, the disorder increases. Entropy is a measure of disorder and water molecules in an ice cube have lower, water molecules in steam have higher entropy.

Ludwig Boltzmann (1844–1906) has changed the entropy topic in science and he was the first to succeed in interpreting entropy from a microscopic point of view (Dugdale, 1998: 86). He relates the entropy to the number of states of a system or arrangements of a system consisting of a large number of atoms and his formulation of the Second Law was based on probability. The concept of entropy has been expressed by Boltzmann in terms of probability of various states. In the year of 1904 Gibbs tried to calculate macroscopic properties of a huge number of particles and explained his idea in a book entitled, “Elementary Principles in Statistical Mechanics”. This book presents the basis for “Statistical mechanics”.

Entropy and irreversibility concepts have a close connection. For example air leaks out of a bursting balloon, but never leaks back in. Hot drinks cool down to room temperature, but never heat up. Arrow of time goes one way. For Planck, irreversible process is essential for the definition of the entropy. “The problem of irreversibility or how to explain the origin of irreversible processes becomes a subject of an intense debate among scientists in the second part of the 19th century and Ludwig Boltzmann made important contribution to this subject” (Benguigui, 2013: 308).

In our study, explaining entropy as a measure of information issue takes place as a second section after an introduction. Relative entropy, Kullback-Leibler divergence concepts within the framework of Bayesian statistics have been explained in the third section. In the fourth
section, measurement of "surprise" by the concept of relative entropy will be studied. The study is completed by a results and discussion section.

2. ENTROPY AS A MEASURE OF INFORMATION

There is always controversy on the meaning of “entropy”:

“Entropy has been characterized variously and vaguely by the words decadence, decay, anarchy, dystopia, randomness, chaos, decay, waste, inefficiency, dissipation, loss of available energy, irreversibility, indeterminacy, most probable states, equilibrium, thermal death, the cause of or the companion of the unidirectional arrow of time. In the social sphere it has been characterized as apocalypse, disorder, disorganization, disappearance of distinctions, meaninglessness, absurdity, uncertainty, pandemonium, loss of information, inert uniformity, and incoherence. The situation is complex because most of these words carry multiple meanings” (Davis, 2011: 3).

A statistical concept of entropy was introduced in 1948 by Claude Shannon (1916–2001) while he was working at Bell Laboratories. At that time his seminal work, “A Mathematical Theory of Communication” was published. At first Shannon could not make a choice between “information” and “uncertainty” and he discussed it with John von Neumann. “Entropy” was the Von Neumann’s choice. “The basic idea, here, is that the amount of information transmitted by a message can be measured. One can then compare the performance of different transmission devices of different coding systems so as to minimize the duration of a transmission or the size of a memory” (Balian, 2003: 122).

Shannon’s entropy measure for a random variable X with values in a finite set can be calculated as follows and negative sign in the formula assures that information measure is positive:

\[ H(X) = - \sum_{i=1}^{n} p_i(x) \log_2 p_i(x) \]

As can be seen, \( H(X) \) is an expected value and can be interpreted as an average in bits. If a random variable has equally likely 8 possible states, we would need 3 bits to transmit the message:
H(x) = - \sum_{i=1}^{n} \left( \frac{1}{B} \right) \log_2 \left( \frac{1}{B} \right) = 3 \text{ bits}

For a non-uniform distribution we can calculate entropy with R:

```r
> p1 = c(0.5, 0.4, 0.10)
> - (0.5*stock(0.5,2)+0.4*stock(0.4,2)+0.10*stock(0.19,2))
[1] 1.268364 \text{ bits}
```

If \( X \) is a continuous random variable with density function \( f(x) \):

\[
H(X) = - \int_{-\infty}^{\infty} f(x) \log f(x) \, dx
\]

For two normal densities entropies and histograms can be calculated and drawn with R as follows:

```r
> x1 = rnorm(10000,0,1)
> y1 = discretize(x1, numBins=10, r=c(0,1))
> entropy(y1)
[1] 2.291397
> hist(x1)
```

![Histogram of x1](image)

```r
> x2 = rnorm(10000,0,10)
> y2 = discretize(x2, numBins=10, r=c(0,1))
> entropy(y2)
[1] 2.29323
> hist(x2)
```
When one event is certain i.e. has a probability of one while all other events have zero probabilities $H(X)$ will have its minimum and be equal to zero. If number of events is denoted by $n$, $H(X)$ reaches to its maximum where all probabilities are equal to $1/n$ and this is the most random situation. Non-uniform distributions have smaller entropy than uniform distributions. If a distribution has a smaller variance (peaked distribution) it has low entropy, if it has larger variance (flat distribution) it has high entropy.

Without a probability distribution it is impossible to calculate the entropy of this distribution. Shannon’s entropy has a partly subjective character, because it characterizes the uncertainty of observers and such a concept fits subjective probability distributions in a Bayesian context (Balian, 2003: 127). The amount $- \log_2 p(x)$ can be considered as a measure of the information brought by the observation of $x$ knowing the probability distribution $p$. Also this amount is the surprise in observing $x$, given prior knowledge on the source summarized in $p$ (Lesne, 2011: 4). Here we may conclude that the concept of “surprise” is also related to Shannon’s entropy measure. Shannon information and its statistical average, Shannon entropy $H(X)$ also measures the amount of uncertainty represented by a subjective probability distribution.
3. RELATIVE ENTROPY: KULLBACK LEIBLER DIVERGENCE

The concept of relative entropy, also is known as “relative information”, “information gain” or “Kullback-Leibler divergence”. If we have two probability distributions such as p and q on the same finite set X, we define the information of p relative to q as:

$$\text{KL}(p \text{ II } q) = \int p(x) \ln \left( \frac{p(x)}{q(x)} \right) \quad \text{KL}(p \text{ II } q) = \sum p(x) \ln \left( \frac{p(x)}{q(x)} \right)$$

Note that relative entropy is not a symmetrical quantity: KL(p II q) ≠ KL(q II p) and takes values in $[0, \infty]$ (Baez and Fritz, 2014: 1).

“Intuitively speaking, $S(q, p)$ is the expected amount of information gained when we discover the probability distribution is really $q$, when we had thought it was $p$. We should think of $p$ as a ‘prior’. When we take $p$ to be the uniform distribution on $X$, relative entropy reduces to the ordinary Shannon entropy, up to a sign and an additive constant. The advantage of relative entropy is that it makes the role of the prior explicit” (Baez and Fritz, 2014: 2).

What is the difference between Shannon’s entropy and relative entropy? The relative entropy measures both the signal and dispersion parts of the information content from observations, while the Shannon’s entropy difference measures only the dispersion part (Xu, 2007: 198).

We can calculate Kullback-Leibler divergence using R such as this (http://cran.r-project.org/web/packages/entropy/entropy.pdf):

```r
> library("entropy")
> # probabilities for two random variables
> freqs1 = c(1/5, 1/5, 3/5)
> freqs2 = c(1/10, 4/10, 1/2)
> # KL divergence from X1 to X2
> KL.plugin(freqs1, freqs2)
[1] 0.1093929
> # and corresponding (half) chi-squared statistic
> 0.5*chi2.plugin(freqs1, freqs2)
[1] 0.11
```

4. HOW TO MEASURE SURPRISE USING ENTROPY

Wikipedia defines “surprise” as, “a brief emotional state experienced as the result of an unexpected significant event” (http://en.wikipedia.org/wiki/Surprise). Surprise can be accepted as an information-theoretic concept. Even surprise and expectations can be used in analyzing music pieces:
“...to a large degree, the way to understand the effect of music is to focus on this ‘kinetics’ of expectation and surprise. The business of making predictions and assessing surprise is essentially one of reasoning under conditions of uncertainty and manipulating degrees of belief about the various proposition which may or may not hold” (Abdallah and Plumbley 2009: 1).

Psychologist Rudolf Arnheim has said that, “The less likely an event is to happen, the more information does its occurrence represent” (Arnheim, 1971: 15). Even p-values have been used to measure surprise (Sellke, Bayarri and Berger, 2001: 62). “We use p-values as ‘measures of surprise’: numerical quantifications of the incompatibility of the observed $t$ and the reference distribution” (Bayarri and Castellanos, 2007: 366). Of course in statistics, the simple outlier detection is also can be regarded as finding a surprise. But surprise index differs from outlier detection by considering an event in the context of other events. Weaver (1948) considers an experiment with n possible outcomes with probabilities $p_1$, $p_2$, ..., $p_n$ by using a “surprise index” and the SI measures “whether the probability realized, namely, $p_i$ is small as compared with the probability that one can accept on the average to realize, namely, $E(p)$ (Baldi and Itti, 2010: 654).

$$SI = \frac{E(p)}{p_i} = \frac{\sum p_i^2}{p_i}$$

Let us suppose that we have three types of whether: Sunny, cloudy, rainy and probabilities are respectively 0.10, 0.85, 0.05. If we calculate the surprise indexes for three events:

- $SI_{\text{sunny}} = (0.01+0.7225+0.0025) / 0.10 = 0.735/0.10 = 7.35$
- $SI_{\text{cloudy}} = (0.01+0.7225+0.0025) / 0.85 = 0.735/0.85 = 0.86$
- $SI_{\text{rainy}} = (0.01+0.7225+0.0025) / 0.05 = 0.735/0.05 = 14.7$

It was shown that observing a rainy day with these probabilities has a bigger surprise index.

**Surprise can also be generated by a sudden change in mean or variance of a time series:**

```r
> library(zoo)
> x=c(1,2,4,3,2,3,4,3,7,6,8,1,2,1,3,2,4,2)
> x1=rollapply(x, width = 3, by = 1, FUN = mean)
> plot.ts(x1)
```
The relation between Shannon's I(D) and the Bayesian surprise S(D) is explained by Pierre Baldi, Laurent Itti as follows:

“One simple approach corresponds to outlier detection theory, whereby surprising events are defined as rare events, i.e. events having a low probability. Such a definition of course is closely related to Shannon's theory of information since, by definition, rare events (P(D) small) have high Shannon's information (I(D) = - log P(D) large). While it is easy to see that in many cases, Shannon's I(D) and the Bayesian surprise S(D) are closely related, there exist also specific situations where these two approaches provide clearly distinct answers, and Bayesian surprise matches intuition better” (Baldi and Itti, 2010: 654).

In recent times the question of detecting a surprise is known as outlier detection, saliency detection, novelty detection have been attracting interests from researchers. The framework for Bayesian surprise first proposed by Itti and Baldi (2005). Some others also have focused on the problem of identifying interesting parts of the video (Hendel, Weinshall and Peleg, 2011: 448). Ranganathan and Dellaert defined landmarks as places that yield highly surprising measurements and used Bayesian surprise in computer vision for landmark detection (Ranganathan and Dellaert, 2009: 2018). In this study, surprise was computed for laser scans and the decision of whether a surprise value corresponds to a landmark is made. Bhatia and Chalup (2013) also used KL divergence to measure surprise, their model provides an account
of surprise experienced and mimics human memory by comparing and associating new patterns with existing patterns and features. Besides for them the concept of surprise can also be understood through its relationship to memory and something that has not been observed stimulates surprise (Bhatia and Chalup, 2013: 140). Some researchers found that surprising events provoke animals to learn faster (Courville, Daw and Touretzky, 2006: 294). Also Kullback-Leibler divergence is used for facial expression recognition and similarity of images is calculated by KL divergence (Huang and Tai, 2012: 45). Besides we have finding that surprising events provoke animals to learn faster (Courville, Daw and Touretzky, 2006: 294).

Surprise can be best quantified by probabilistic framework of Bayesian theory. Surprise is subjective and a rare piece of data with small probability is considered more informative (Baldi and Itti, 2010: 649); for the same visual data surprise degree differs for different individuals. In Bayesian surprise theory, we have a prior probability distribution and when the new stimuli arrive the prior probability distribution is updated and transformed into a posterior probability distribution. In such a case, we measure surprise by KL divergence between prior and posterior distributions. Larger KL divergence leads larger surprise. If data do not affect prior probabilities then there is no surprise.

“Starting from Bayes theorem, we notice that the fundamental effect that data has on a given observer is to change his/her/its prior beliefs into posterior beliefs. Thus we propose to measure the effect D has on the observer by the distance between his prior and posterior belief distributions. We call this facet of information surprise” (Baldi and Itti, 2010: 649).

It is clear that surprising events that occur repeatedly will cease to be surprising and events with very low probability will be regarded as surprising only if they induce a meaningful change in the observer's beliefs (Hendel, Weinshall and Peleg, 2011: 454). Process of surprise with prior and posterior distributions is shown in Figure 1.
Pierre Baldi explains how relative entropy measures surprise within the context of Bayesian theory:

“Bayesian surprise quantifies how data affects natural or artificial observers, by measuring differences between posterior and prior beliefs of the observers. Using this framework we tested whether humans orient their gaze towards surprising events or items while watching television. Bayesian surprise strongly attracts human observers, with 72% of all gaze shifts directed towards locations more surprising than the average, a figure rising to 84% when considering only gaze targets simultaneously selected by all subjects” (http://ilab.usc.edu/surprise/).

The dissimilarity between two probability distributions is measured by relative entropy. Therefore, the relative entropy is also called as "distance", Kullback-Leibler divergence (relative entropy) for two probability distributions as defined as

\[ \text{KL}(p \parallel q) = \sum p(x) \ln \left( \frac{p(x)}{q(x)} \right) \]

The use of Bayesian statistics in the context of relative entropy can be explained in the following manner. As is known, Bayes theorem states how a prior distribution of \( p(x) \) transforms into a \( q(x) \) posterior distribution when a new \( Y = y \) data is observed,

*Entropy of the prior distribution: \( H(P(X)) = -\sum p(x) \log p(x) \)*

*Entropy of the posterior distribution: \( H(P(X \mid Y)) = -\sum p(x \mid y) \log p(x \mid y) \)*

The information gain by observing \( y \), can be calculated as the Kullback-Liebler divergence:

\[ \text{KL}(p \parallel q) = \sum p(x) \ln \left( \frac{p(x)}{q(x)} \right) \]
5. RESULTS AND DISCUSSION

Surprise is a subjective emotion and it is related to rare events. Outliers, p-values, surprise indexes are statistical measures for surprise. Surprise attracts human attention and measurement of this concept can be useful for various applications. In the context of Bayesian theory, surprise can be measured statistically using relative entropy and in such cases, Pierre Baldi and Laurent Itti proposed Kullback-Leibler divergence to measure surprise. If we find the dissimilarity between prior and posterior distributions in Bayesian context this quantity is supposed to measure surprise.
REFERENCES

• Package ‘entropy’, August 29 2013,
• http://cran.r-project.org/web/packages/entropy/entropy.pdf, 28 Feb 2014.
• http://ilab.usc.edu/surprise/, 15 Feb 2014.