Testing The Torah Code Hypothesis: The Experimental Protocol

Robert M. Haralick
Computer Science, Graduate Center
City University of New York
365 Fifth Avenue
New York, NY 10016

Abstract

This is the second part of a tutorial discussing the experimental protocol issues in Testing the Torah Code Hypothesis. The principal concept is the test statistic which is used to do the actual hypothesis testing of the Null hypothesis against a simple alternative or against a complex of alternatives.

We illustrate the methodology using the data sets from the WRR[3] experiment. We use the WRR key word sets of list 1 and 2 combined. The experiment produces a p-value of less than 1/100,000 in the Genesis text.

We performed another experiment pairing rule based transliterations for the spellings of the names of the American presidents into Hebrew with the Hebrew word for president. Taking into account Bonferroni, the resulting p-value of the 100,000 trial experiment was less than 1/66,667.

1. Introduction

In order to have an experiment that is reproducible, there has to be an experimental protocol which describes in sufficiently precise detail all the steps and calculations so that another researcher can independently perform the experiment and expect to get results that are insignificantly different from that of the original experiment. It is this kind of replication that the scientific methodology demands. In this paper we provide exact descriptions of experimental protocols that can test different variations of the Torah Code hypothesis. The notation and concepts in this paper follow that of Haralick[1] and we do not repeat here any of the definitions or concepts discussed there.

Just as in pattern recognition, where it is well known that some features will work better for a particular task so in testing the Torah code hypothesis, some protocols work better than others, better meaning lower false alarm and misdetection rates. At this time it is not known what the best protocol is, but in this paper we are able to demonstrate a protocol with an improved false alarm rate compared to the original WRR protocol.

Principal concepts involve the control population, here called the monkey text population and the test statistic for actually doing the hypothesis testing of the Null hypothesis against a simple alternative and against a complex of alternatives each associated with the Torah Code hypothesis. Our test statistic uses multiple compactness features and its formula is motivated by a probability derivation. Our experimental protocol uses the test statistic as a score in Monte Carlo experiment

We first illustrate the application of the experimental protocol in an experiment of the McKay key word set in the *War and Peace* text and the WRR key word set of list 1 and list 2 combined in the *Genesis* text. The McKay experiments were an attempt to illustrate that the codes found in the Torah text could be replicated in an ordinary text by sufficient wiggling and fiddling with the key word choices and spellings. With our protocol, we are not able to reject the Null hypothesis for the McKay key word set in the *War and Peace* text and we must reject the Null hypothesis for the WRR key word set in the *Genesis* text. Then we illustrate the application of the experimental protocol in an experiment originally suggested by 13 year old David Roffman in December 2005: the relationship between the names of the American presidents and the Hebrew word **נשיא**, meaning president.

2. Hypothesis Testing

The formal way in which the significance of an encoding is evaluated is by a test of Hypotheses. The Null hypothesis of No Torah Code Effect is tested against an alternative hypothesis that there is an encoding.

The statistical computation involved in the test of hypotheses amounts to determining the fraction of monkey texts that have at least as good an encoding as the Torah text. Or saying it another way, if the compactness value of the given key word set in the Torah text is v_1 and the com-

0-7695-2521-0/06/$20.00 (c) 2006 IEEE

pactness value of the given key word set in Monkey texts $2, ..., N$ is $v_2, \ldots, v_N$ then the estimated probability that a monkey text would have as good an encoding as the Torah text is the normalized rank of v_1 among $v_1, \ldots, v_N$. This normalized rank is called the p-value of the experiment.

To do the test of hypothesis, the p-value of the experiment is compared to a significance level α_0. If the p-value is smaller than α_0, then the Null hypothesis of No Torah Code effect is rejected in favor of the alternative that the key word set has ELSs in an unusually compact arrangement. If the p-value is larger than the significane level, the Null hypothesis is not rejected.

2.1. Test of Null Hypothesis Against A Complex Alternative Hypothesis

An experiment about a particular historical event is described by a set of what are considered to be the key words relevant to that event. However, not all of the key words thought about might have corresponding ELSs in a relatively compact arrangement. Hence an hypothesis test of the Null hypothesis against the Alternative hypothesis that all of the key words have ELSs that are in a relatively compact arrangement will most likely not be rejected.

Therefore, the experiment is set up as a test of the Null hypothesis against multiple alternatives hypotheses. Each alternative hypothesis is specified by some subset of the given total set of key words. The formal test of the Null hypothesis is against the alternative hypothesis that at least one of the alternative hypotheses is true.

2.2. Bonferroni

When K separate experiments are done, each testing the Null hypothesis against a different Alternative hypothesis, yielding p-values $p_1, \ldots, p_K$, the smallest p-value is not the p-value of the complex of the K separate experiments. Indeed, if the experiments are separate, then the exact p-value of the complex of K separate experiments cannot be determined if they are not independent. This is the usual case. However, it can be bounded. The Bonferroni upper bound is $K \min\{p_1, \ldots, p_K\}$. The p-value of the K separate experiments must be smaller than the Bonferroni bound. Therefore, if the Bonferroni bound which is necessarily higher than the p-value of the complex of experiments, is smaller than the significance level, then it necessarily follows that the p-value of the complex of experiments is also smaller than the significance level. In this case the Null hypothesis can be rejected at the given signficance level.

The problem with the Bonferonni bound is that it is an upper bound and in many instances is much higher than the true p-value of the complex of K separate experiments. This is particularly true when the K Alternative hypothesis are statistically dependent.

2.3. K Scores And Combine

There is a statistically economical alternative to using the Bonferroni bound when testing the Null hypothesis against a complex of K Alternative hypotheses. The alternative is on a trial by trial basis, to use K scoring schemes, one appropriate for each of the K Alternative hypotheses, and then combine the scores together in a suitable way.

Suppose there are K key word sets, each describing the same historical event. In this case it is expected that every pair of key word sets will have a substantial fraction of its key words in common to both sets. In this case the Alternative hypotheses will necessarily have statistical dependence.

Each trial of an N trial experiment randomly samples a monkey text from the monkey text population. In accordance with a specified protocol, on trial n, the compactness of the ELSs from each of the K key word sets is computed, resulting in $c_{1n}, \ldots, c_{Kn}$. For the k^{th} key word set, the compactness values $c_{k1}, \ldots, c_{kN}$ of the N trials are rank normalized to $r_{k1}, \ldots, r_{kN}$.

The p-value associated with test of the Null hypothesis against the Alternative that the k^{th} key word set has its ELSs in a more compact relationship than expected by chance is given by r_{k1}. The Bonferroni bound B on the test of the Null hypothesis against the K alternative hypotheses is then $B = K \min\{r_{11}, \ldots, r_{K1}\}$.

The K scores and combine feature, would define a combining function F acting on the rank normalized values $r_{1n}, \ldots, r_{Kn}$, for the n^{th} trial of the experiment. In this situation, combining functions ought to be symmetric in its arguments. For example, one combining function could be the minimum: $f_n = F(r_{1n}, \ldots, r_{Kn}) = \min\{r_{1n}, \ldots, r_{Kn}\}$. The scores $f_1, \ldots, f_N$ are rank normalized and the rank normalized value, p_1, associated with f_1 is the p-value of the experiment. For the min combining function, p_1 is nessarily smaller than the Bonferonni bound B.

However, the min combining function is not necessarily the statistically most optimal. For example, a combining function may be motivated by a probability derivation that has even some unwarranted conditional independence assumptions.[1] One such combining function is

$$\begin{aligned} F_1(r_{1n}, \ldots, r_{Kn}; \theta) &= \frac{1}{K} \sum_{k=1}^{K} p(r_{kn}; \theta) \\ &= f_{1n} \end{aligned} \quad (1)$$

[1] The unwarranted assumptions are not used in making any probability calculations for p-value. The probability derived by using the unwarranted assumptions gives a motivation and a formula for performing a calculation of a score function. It is the score function that is used in a proper Monte Carlo experiment for determining the p-values.

where $p(r;\theta)$ is the probability under the Alternative hypothesis of observing a normalized relative rank of r in an N trial experiment and $1/K$ is the prior probability of any one of the K alternative hypotheses of being true. We base $p(r;\theta)$ on $-\log$ because for small relative ranks $-\log$ will be large ($\log(2N)$) for an N trial experiment where the smallest and unique relative rank is $1/2N$. For larger relative ranks, $-\log$ will be small and indeed be 0 for a relative rank of 1.

$$p(r;\theta) = \begin{cases} -\beta(\theta)\log(r) & \text{when } r < \theta \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

This combining function arises (up to a fixed constant of proportionality) when exactly one of the K alternative hypothesis is assumed to be true, and for each trial n, and for each k, the random variables $r_{1n}, \ldots, r_{Kn}$ are conditionally independent given that Alternative Hypothesis k is true. When the Null hypothesis holds the remaining $K-1$ possibilities are assumed to follow the discrete uniform on the normalized relative ranks of an N trial experiment. The threshold θ specifies that the probability of observing a normalized rank greater than θ under the alternative hypothesis is 0. We call this method the *first order combining method*.

If two of the K alternative hypotheses are assumed to be true, with the prior probability for any pair to be true to be $2/(K(K-1))$, then under the same conditions as the previous derivation, the combining function should be

$$F_2(r_{1n}, \ldots, r_{Kn}; \theta) = \frac{2}{K(K-1)} \sum_{\{(j,k)|k>j\}} p(r_{jn};\theta)p(r_{kn};\theta)$$

$$= f_{2n} \quad (3)$$

We call this method the *second order combining method.*

If it is assumed that when there is an encoding either one or two of the K alternative hypotheses is true, and the prior probability for exactly one alternative being encoded is the same as the prior probability for exactly two alternatives being encoded, then under the same conditions as the first probability derivation, the score s_n of the n^{th} trial should be

$$s_n = f_{1n}/N + f_{2n} \quad (4)$$

We call this method the *non-rank normalized* method of combining.

Another possible way of combining F_1 with F_2 would be to take the N values $f_{11}, \ldots, f_{1N}$ and rank normalize them forming the N normalized ranks $t_{11}, \ldots, t_{1N}$. Also rank normalize the N values $f_{21}, \ldots, f_{2N}$ forming the N normalized ranks $t_{21}, \ldots, t_{2N}$. Define the rank normalized score s_n for the n^{th} trial by the convex combination

$$s_n = wt_{1n} + (1-w)t_{2n} \quad (5)$$

for a specified weight w.

In either combining method, the p-value of the experiment is the normalized relative rank of the score for the first trial.

2.4. Composite Experiments

Composite experiments are associated with multiple events. Suppose that there are M events. Each event has associated with it a collection of key word sets. The m^{th} such collection is associated with a test of the Null hypothesis against the K_m alternative hypotheses formed by each one of the K_m key word sets in the collection. In the composite experiment, we are interested in a test of hypotheses at two levels. First we wish to test the Null hypothesis against the Alternative that more of the M events have their ELSs in a more compact arrangement than expected by chance. Second we wish to test the Null hypothesis against the Alternative that more of the $K_1 + \ldots + K_M$ key word sets have their ELSs in a more compact arrangement than expect by chance.

In the first case, we treat each event as an experiment that produces in each trial a score which is the normalized relative rank of the compactness associated with the trial. Thus each trial produces M scores. These scores then combined together in a test statistic appropriate for a test of the Null hypothesis against M Alternative hypotheses, where one of the M Alternative hypotheses is assumed to be true. We know that in this situation, the probability of small scores under the Alternative hypothesis for each of the M alternatives is not as high as in the case when considering the probability of small scores of an alternative in the complex of alternative situation. In this case, suitable combining function choices include

$$g_n = G(s_{1n}, \ldots, s_{Mn}) = \frac{\beta}{M} \sum_{m=1}^{M} 1 - \exp(-s_{mn}/q_m) \quad (6)$$

or

$$g_n = G(s_{1n}, \ldots, s_{Mn}) = \frac{\beta}{M} \sum_{m=1}^{M} \log(1. - s_{mn}/q_m) \quad (7)$$

where s_{mn} is the score of the m^{th} event of the n^{th} trial and q_m is a scale factor computed as the largest of the scores for the m^{th} event.

$$q_m = \max_{n=1,\ldots,N} s_{mn}$$

The p-value of the composite experiment is the normalized relative rank of g_1.

3. Compactness Measures

It has been anecdotally noticed that sometimes ELSs have compact meetings in accordance with one compactness measure and other times in accordance with a different compactness measure. Therefore an experiment can select multiple kinds of measures that tend to score well for many of the kinds of compact meetings noticed. These different compactness measures are in effect associated with different alternative hypotheses of Torah code effect.

The raw rank normalized data set of the M collections of key word sets, one collection per event, for the n^{th} trial then can be represented by

$$r^c_{11n}, \ldots, r^c_{1K_1n}; r^c_{21n}, \ldots, r^c_{2K_2n}; \ldots; r^c_{M1n}, \ldots, r^c_{MK_Mn}$$

$c = 1, \ldots, C; n = 1, \ldots, N$ where C is the number of compactness measures, N is the number of trials, and K_m is the number of key word sets in the collection of key words sets associated with the m^{th} event.

In accordance with either the non-rank normalized method or the rank normalized method, the raw data set is processed to produce scores for each event, trial, and compactness measure. We denote by s^c_{mn} the score associated with event m, trial n, and compactness measure c.

If the alternative hypothesis is that the encoding occurs with at least one of the compactness measures, then it is reasonable to form a score s_{mn} by

$$s_{mn} = \max_{c=1,\ldots,C} s^c_{mn} \tag{8}$$

If the alternative hypothesis is that each encoding occurs with all of of the compactness measures simultaneously, then it is reasonable to form a score s_{mn} by

$$s_{mn} = \min_{c=1,\ldots,C} s^c_{mn} \tag{9}$$

4. Our Experimental Protocol

We use the combined data from list 1 and list 2 of the WRR paper[3]. This data set has become a standard in Torah code work, similar to the status of how the Fisher Iris data set is used in the classic discrimination experiments. For the WRR data, each key word set has one appellation and one date of either death or birth. There are 53 rabbinic personalities for which at least one of its key word sets has at least one ELS for each of the key words in the set. These rabbinic personalities represent our events. The collection of key word sets associated with each rabbinic personality constitute the possible event descriptions. There are a total of 321 key word sets for these 53 rabbinic personalities.

For our skip specification σ, we set the largest skip permitted for ELSs of a given key word to be such that the expected number of ELSs searching from a minimum skip of 2 would be 10. This is similar to the protocol of WRR. And we set the minimum skip for ELSs to be 1 (WRR sets the minimum skip to be 2).

For our resonance specification ϕ, we require that at least one ELS from each key word in a key word set be resonant on a cylinder size and on the resonant cylinder size the skip of the ELS must be no more than 10 rows and no more than 10 columns. This differs from WRR who insisted that for one ELS the row skip on the cylinder be no more than 10 rows.

For our monkey text population we use the ELS random placement population with 100,000 trials. The Monte Carlo experiment is carried out with an independent execution for each rabbi. The random number seed was obtained from the digits of π. Starting from the first digit after the decimal point, the digits were broken up into strings of seven long. Each successive string of seven π digits was used as the random number seed for each successive rabbi Monte Carlo experiment.

We use both first order (1) and second order (3) methods of combining over the key word sets of each rabbi. We set our threshold $\theta = .2$ in (2), a value used by WRR[3] in a slightly different context, but in the same spirit as we used it.

For our compactness features we choose two kinds of compactnesses that measure essentially different kinds of geometric arrangement. The *1D* compactness measure searches over all ELS sets ζ satisfying the skip specification, and finds the ELS set having the smallest span length and the corresponding text segment. The *span length* of an ELS set is the difference between the largest ending position taken over all ELSs in the set and the smallest beginning position taken over all ELSs in the set. This compactness feature is essentially the area of the table formed on a cylinder of 1 column. Our second compactness feature is distance based. Following the notation in Haralick[1], it is formed by R_{12} followed by Ψ_{harm}.

For each of the 100,000 trials, each of the 321 key word sets[2] has its ELSs evaluated by the two compactness measures. Each of the resulting 642 compactness values are then rank normalized producing the raw rank normalized data for performing the hypothesis testing.

On each trial, for each rabbinic personality, for each of the two compactness measures, we combine using (1) and (3) over the key word sets associated with the rabbinic personality. We use the normalized rank method (5) of combining the first order and second order ranks together. As in (9) the minimum of the resulting two compactness values is then the score for the trial and rabbinic personality.

On each of the 100,000 trials, these 53 scores are com-

[2]We are only counting those which have at least one ELS for each key word in the set.

bined using the g combining method using the exp function (6). The relative rank of the g-score for the first trial is the p-value associated with the test of the Null hypothesis against the alternative that more of the 53 rabbinic personalities have ELSs from one or two of its key word sets in a more compact arrangement by both the 1D and the distance compactness measure than expected by chance. The p-value of this experiment was $.5 \times 10^{-5}$. For reference purposes the p-value of an identical experiment using the WRR list 1 was 2.75×10^{-3} and for WRR list 2 was 2.5×10^{-4}. None of the WRR lists produced a significant p-value on the Hebrew translation of the *War and Peace* text.

5. The McKay Demonstration

McKay[2] tried to illustrate that the successful experiment of WRR[3] could be "re-enacted" in the Hebrew translation of *War and Peace* if one fiddled and wiggled enough in making changes in spellings (including incorrect spellings) and choices of appellations (including incorrect appelations). Indeed, their demonstration in the *War and Peace* text produced a p-value of about 1/1,000,000 using the same protocol as WRR. Their conclusion was that the success of the WRR experiment was due to choice in the input data of appellations and dates and not due to a genuine ELS phenomena in *Genesis*. There is no space here to explain the various technical problems with the McKay et. al. paper. We just want to note that the McKay data set of appellations for *War and Peace* produces a p-value of .06585 in an experiment of 10,000 trials with exactly the same experimental protocol as employed in our experiment in the *Genesis* text. Clearly, the McKay data set does not produce a significant p-value in the *Genesis* text.

There is an important interpretation that one can make from these results: there is a structural/geometric difference between the ELS arrangements of the McKay appellations in *War and Peace*, which do not constitute any encoding, from that of the WRR appellations in *Genesis*, which are hypothesized to be an encoding. It follows from this result that the protocol used by WRR and McKay was not sensitive enough to detect this geometric difference. Or saying this another way, the inherent false alarm rate with the WRR protocol is higher than with our protocol. That was the reason McKay was able to make his demonstration succeed.

6. The American Presidents

In this section we report on an experiment pairing the names of the American presidents, transliterated into Hebrew with the key word נשיא, meaning *president*. There are 42 people who have served as presidents, some multiple times. Due to the various ways non-Hebraic names can be spelled in Hebrew, we devised a rule base system to provide a reasonable set of Hebrew spellings for each president's name. The rule base is described in the appendix. As an example, the name Lincoln is transliterated as לינקולן, לנקולן, לינקלן, or לנקלן. In addition we use two variations: the last name alone and the first character of the president's first name as a prefix to the spelling of the last name. The total number of spellings having ELSs was 248, on the average nearly six spellings per name. The p-value using the identical protocol as in section 4 with 1000 trials was not significant.

We performed a second experiment using just the compactness measure defined by R_{12} followed by ψ_{min} followed by Ψ_{harm}. The p-value was .0045. This indicated that something interesting was happening. So we explored further. We examined the distance measure formed by Ω_2 followed by ψ_{min} followed by Ψ_{harm} in a 100,000 trial experiment. This is a compactness measure reported on at the International Torah code conference a few years ago. With this compactness measure and our protocol, the American president experiment tests the Null hypothesis of *No Torah Code Effect* against the complex alternative hypothesis that

1. in accordance with the Hebrew to English transliteration rules of the appendix (section 8)
2. and in accordance with the skip specification, and the resonance specification stated in section 4
3. for nearly all the presidents
4. each president has one or two Hebrew spellings of his name
5. that have ELSs which are in a more compact arrangement with ELSs of the Hebrew word נשיא, meaning president
6. in the 5 books of the Chumash
7. by compactness measure Ω_2 followed by ψ_{min} followed by Ψ_{harm}

In a 100,000 trial experiment, the resulting p-value was .000005, the smallest p-value possible. Clearly, the experiment has to be repeated with more trials to get a better estimate of the p-value. At this point we have done three experiments. By Bonferroni, we can only bound the true p-value to be less than 3/200,000=1/66,667.

7. Concluding Discussion

We have discussed experimental protocol possibilities by which an experiment can be done that tests the Null hypothesis of No Torah Code Effect against a composite alternative. The composite alternative is that more of the 53

rabbinic personalities have ELSs of their key word sets in a more compact arrangement by both a 1D compactness measure and a distance compactness measure than expected by chance. For this purpose we developed a score function based on a probability derivation of what the probability would be if one or if one or two of a fixed number of choices follows a given probability function while the remaining follow a discrete uniform probability function.

For various reasons, that we did not discuss due to space limitations, our methodology is more conservative than that employed by WRR[3]. We performed a 100,000 trial experiment that took more than 36 hours on an AMD 64 X2 4400 processor. The experiment produced a p-value of .5/100,000. It is clear that the Null hypothesis of No Torah Code Effect has to be rejected. The resulting p-values were so small that a 15 day experiment of 1,000,000 trials needs to be done to get better estimates of how small they really are.

The protocol used in this experiment was developed (trained on) the WRR data set. The protocol is direct, statistically motivated, self normalizing, consistent with the nature of the alternative hypothesis, and (in our opinion) aesthetically simple. No part of the protocol has large numbers of variables or parameters whose values can be set to memorize the pattern of the ELS data from the Torah text versus that from the monkey texts. The parameters of the protocol itself were three: maximum skip set so that the expected number of ELSs was about 10; the maximum row and column skip of an ELS on a cylinder was 10. The probability threshold was .2. The rest of the freedom in the protocol came from methodological choices: the monkey text population, the compactness measures, the various rank normalizations, the combining method over key word sets of an event, the combining method over scores of events.

For our future work, we will be applying this protocol to new data sets.

8. Appendix: Transliteration of English Names Into Hebrew

Here we give the principles by which the English names of the presidents were transliterated into Hebrew in all the possible forms. The transliteration of the consonants are shown in the first table and the transliteration of the long and short vowels are shown in the second table.

The rule we used is to transliterate each name with every combination of the vowels formed by keeping or omitting the Hebrew wowel. Hence a name with two vowels will have four possible transliterations.

B	ב	P	פ
C (see)	צ	Q	ק
C, ck (kay)	ק	R	ר
D	ד	S (ess)	ס
F	פ	S (zee)	ז, ס
G	ג	Sh	ש
H	ה	T	ט
J	ג	Th, Ta	ת
K	ק	V (next to long vowel)	ב
Kn	נ	V (next to short vowel)	ו
L	ל	W	ו
M	מ	X	קס
N	נ	Z	ז

Table of transliteration of English consonants into Hebrew consonants

English Vowel	Long Vowel	Hebrew	Short Vowel	Hebrew
	C**a**ke	א	c**a**t	-
A	H**a**yes,	״	Buch**a**nan	א
	T**ay**lor	״	**A**dams,	א
	R**ea**gen	״	**A**rthur	א
			T**a**ft,	א
			Gr**a**nt	א
			C**a**rter	א
E	s**ee**k, b**ea**d	י	s**e**t	-
	f**ie**ld, P**ie**rce	י	J**e**fferson	-
	b**i**ke	אי, יי	F**i**llmore	י
I	T**y**ler	אי, יי	Mad**i**son	י
			Cl**i**nton	י
			Harr**i**son	י
			N**i**xon	י
			W**i**lson	י
	b**oa**t, r**o**se	ו	Wils**o**n,	ו
	R**oo**sevelt	ו	Clint**o**n	ו
O	P**o**lk	ו		
U	Tr**u**man,	ו	p**u**p,	ו
	H**oo**ver	ו	Roos**e**velt	ו

Table of transliteration of English vowels into Hebrew consonants

References

[1] R. Haralick. Basic concepts for testing the torah code hypothesis. In *ICPR*, 2006.

[2] B. McKay, D. Bar-Natan, M. Bar-Hillel, and G. Kalai. Solving the bible codes puzzle. *Statistical Science*, pages 149–173, May 1999.

[3] D. Witztum, E. Rips, and Y. Rosenberg. Equidistant letter sequences in the book of Genesis. *Statistical Science*, 9(3):429–438, August 1994.