Finding Uncertainty Reduction in Consecutive Prime Residues with Reconstructability Analysis

Shawn Marincas
Systems Science PhD Student

Portland State University

December 6, 2019

Why?

- Data Mining with Information Theory (SYSC 431/531).
- Lemke Oliver and Soundararajan paper "Unexpected Biases in the Distribution of Consecutive Primes" [1] (2016).
- Prime numbers and reconstructability analysis (RA) are neat.
- Can we find patterns in consecutive prime numbers using RA?

Who?

Shawn Marincas, a systems science PhD student and professional programmer and tech consultant.

Not a mathematician.

What?

Data Models

If you know the full probability distribution of every variable together for a set of unsorted categorical data, then you would be able to recreate, with maximum accuracy, the original unsorted categorical data.

This probability distribution is the "saturated" model and represents the most accurate possible model and the highest possible complexity for the data.

Model Complexity

However, if your data is of variables which are independent of each other, then you would be able to recreate the data just as accurately, with a less complex model.

We want to find the least complex model which will most accurately recreate the original dataset. This model defines our best estimation of the "structure" that exists within that dataset.

Information Theory

Information Entropy (Shannon 1949) is the negative logarithm of the probability mass function for a set of states.

$$H(x) = -\sum_{j}^{n} p_j log_2 p_j \tag{1}$$

For a set of categorical data, this is a function of the probability distributions of each variable, or set of variables, and describes how diverse the distribution is. A uniform probability distribution has the highest entropy, so the higher the entropy, the more uncertain the model is.

Model Uncertainty

These sets of variables are the models of knowledge, and entropy is a measure of diversity, or uncertainty, in that model.

The independence model will have the highest entropy, and the saturated model will have the lowest.

Data Mining with Information Theory

Compare the information entropy of a model against the entropy of either the full data, or the independence model.

Measuring a model against the independence model tells you how much information that model captures.

Measuring a model against the full data model tells you how much error is in that model.

Prime Numbers

Prime numbers are the numbers which can not be divided evenly by any numbers between 1 and themselves.

Here are all the prime numbers under 100:

2 3 5 7 11 13 17 19 23 29 31 37 41 43 47 53 59 61 67 71 73 79 83 89 97

Consecutive prime numbers would be sequences of neighboring prime numbers on this number line. We will be looking at the pairs of prime which neighbor each other.

Reduced Residue Classes

A residue class ($mod\ q$) is the set of integers which leave a remainder (or residue), r, when divided by another integer, q.

A reduced residue class is one where all r are relatively prime to q, which means that q and r do not share any prime factors.

The important thing here is that a "reduced residue class (mod q)" is identified by the residue r, which is the remainder when dividing by q.

By using q as a categorical variable, the residues, r, can be categorical data about a set of N integers. If those integers are pulled from the set of $\mathbb{P}$ then we have a dataset of prime residues, as seen in Table 2. Notice the values of q are also prime.

Prime Residues

N mod q	3	5	7
11	2	1	4
13	1	3	6
17	2	2	3
19	1	4	5
23	2	3	2

Table: Table of Residues mod $q \epsilon \{3,5,7\}$ of Primes.

Consecutive Prime Residues

We want to look at relationships between consecutive prime numbers using categorical data analysis, specifically reconstructability analysis.

To do this with unsorted categorical data requires us to use "masking" which involves creating lag variables. This means that while the data is technically unsorted, the DV columns contain the values from IV columns in a row "adjacent" to the current row.

This is better explained by example in Table 2 where we see that the Z_α columns representing the DVs contain the values from the corresponding α column in the following the row. By using the Z_α column as our DV we are then looking for models of prime residues which predict the following prime residue.

Consecutive Prime Residue Example

N	N mod 3 A	N mod 5 B	N mod 7 C	$N+1$ mod 3 Z_A	$N+1$ mod 5 Z_C	$N+1$ mod 7 Z_E
11	2	1	4	1	3	6
13	1	3	6	2	2	3
17	2	2	3	1	4	5
19	1	4	5	2	3	2
23	2	3	2	2	4	2

Table: Example of consecutive prime residue data.

Unexpected Biases in the Distribution of Consecutive Primes

Lemke Oliver and Soundararajan found that prime numbers show a pattern in how consecutive reduced residue classes (mod q) of primes would tend not to repeat themselves.

What this means is that given a dataset of including variable (R_0) for reduced residue class (mod q) of a prime, p_n, and a variable (R_1) for the reduced residue class (mod q) of the following prime, p_{n+1}, that a model of R_0 would reduce our uncertainty of R_1.

ID	Model	Level	ΔDF	α	Information	%ΔH(DV)	ΔBIC	Inc. α	Prog.	%C(Data)	%cover
30*	IV:CEGHIZc	5	51837	0	0.98424046	7.6189	20169328.06	0	23	39.5134	100
25*	IV:ACEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	17	39.5029	100
24*	IV:CDEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	18	39.5029	100
23*	IV:CEGHZc	4	4317	0	0.77355558	5.988	16522867.42	0	16	37.889	100
22	IV:ABCEHZc	5	2877	0	0.77349646	5.9876	16548124.28	1	19	37.8883	100
21	IV:BCDEHZc	5	2877	0	0.77349646	5.9876	16548124.28	1	20	37.8883	100
20*	IV:CDEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	12	37.8881	100
19*	IV:ACEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	13	37.8881	100
18*	IV:CDEIZc	4	1725	0	0.74289985	5.7507	15912667.01	0	14	37.583	100
17*	IV:ACEIZc	4	1725	0	0.74289985	5.7507	15912667.01	0	15	37.583	100
16*	IV:CEGZc	3	429	0	0.55980145	4.3334	12006801.98	0	10	36.0308	100
15*	IV:ACEZc	3	141	0	0.55979122	4.3333	12011887.64	0	9	36.03	100
14*	IV:CDEZc	3	141	0	0.55979122	4.3333	12011887.64	0	8	36.03	100
13*	IV:ACHZc	3	237	0	0.49048893	3.7968	10522723.14	0	7	34.8074	100
12*	IV:CDHZc	3	237	0	0.49048893	3.7968	10522723.14	0	7	34.8074	100
11*	IV:CEZc	2	69	0	0.32081986	2.4834	6884305.592	0	5	32.6711	100
10*	IV:CGZc	2	69	0	0.31871201	2.4671	6839065.995	0	4	32.9222	100
9*	IV:ACZc	2	21	0	0.31871053	2.4671	6839918.451	0	3	32.9204	100
8*	IV:CDZc	2	21	0	0.31871053	2.4671	6839918.451	0	2	32.9204	100
7*	IV:CHZc	2	117	0	0.28328243	2.1929	6077776.633	0	6	32.2527	100
6*	IV:CZc	1	9	0	0.17264152	1.3364	3705142.314	0	1	30.4308	100
5*	IV:EZc	1	15	0	0.01267883	0.0981	271842.2082	0	1	26.9375	100
4*	IV:GZc	1	15	0	0.01003313	0.0777	215059.1452	0	1	26.3393	100
3*	IV:AZc	1	3	0	0.01003277	0.0777	215272.5338	0	1	26.3393	100
2*	IV:DZc	1	3	0	0.01003277	0.0777	215272.5338	0	1	26.3393	100

Table: Directed OCCAM Search of loopless models for Z_C, $p_n \pmod 5$)

		observed data				calculated model								
C	freq	Zc=1	Zc=2	Zc=3	Zc=4	Zc=1	Zc=2	Zc=3	Zc=4	rule	#correct	%correct	p(rule)	p(margin)
1	24999432	18.493	30.019	29.718	21.77	18.493	30.019	29.718	21.77	2	7504611	30.019	0	0
2	25000399	25.496	17.757	27.02	29.727	25.496	17.757	27.02	29.727	4	7431869	29.727	0	0
3	25000130	24.044	28.175	17.77	30.011	24.044	28.175	17.77	30.011	4	7502895	30.011	0	0
4	25000024	31.966	24.051	25.492	18.492	31.966	24.051	25.492	18.492	1	7991430	31.966	0	0
	99999985	24.999	25	25	25	24.999	25	25	25	2	30430805	30.431		

Table: OCCAM Fit results for the CZ_C model, the model from Lemke Oliver and Soundaarajan paper.

DV	Model	Level	ΔDF	Alpha	Information	ΔH(DV)	ΔBIC	Inc. Alpha	Progenitor	%C(Data)	% Cover
Zc	IV:CZc	1	9	0	0.17264152	1.3364	3705142.314	0	1	30.4308	100
Zc	IV:CEZc	2	69	0	0.32081986	2.4834	6884305.592	0	5	32.6711	100
Zc	IV:ACEZc	3	141	0	0.55979122	4.3333	12011887.64	0	9	36.03	100
Zc	IV:ACEHZc	4	1437	0	0.773444	5.9872	16573524.21	0	13	37.8881	100
Zc	IV:ACEHIZc	5	17277	0	0.98279785	7.6078	20774984.78	0	17	39.5029	100

Table: Best model per level of a loopless OCCAM directed search for p_n (mod 5), Z_C, up to the best overall model at level 5.

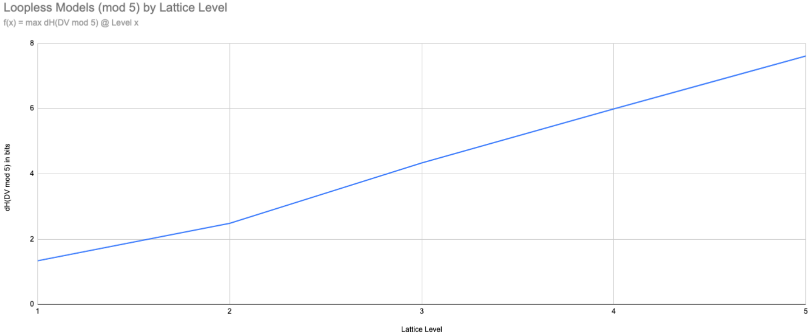

Figure: Best loopless models of uncertainty reduction in p_n (mod 5), Z_C, proceeding up the lattice levels.

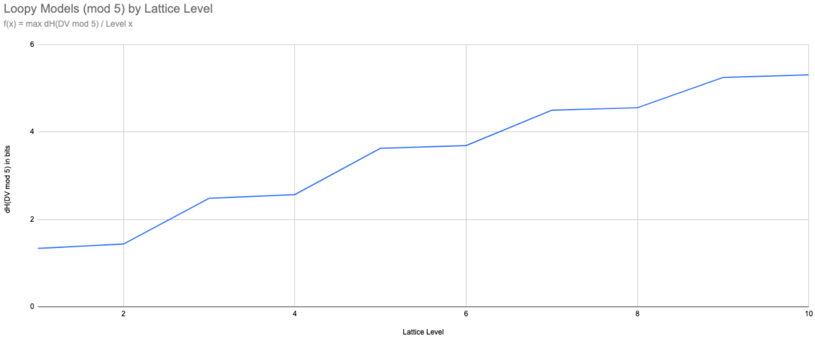

Figure: Best overall models of uncertainty reduction in p_n (mod 5), Z_C, proceeding up the lattice levels.

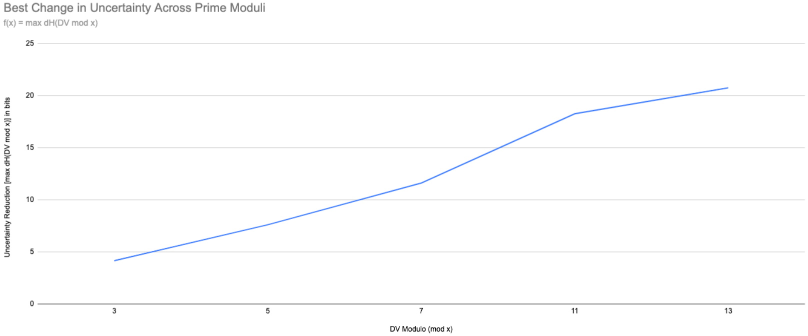

Figure: Best loopless models of uncertainty reduction across DVs.

A	C	E	H	I	freq	Zi=1	Zi=2	Zi=3	Zi=4	Zi=5	Zi=6	Zi=7	Zi=8	Zi=9	Zi=10	Zi=11	Zi=12	rule	#correct	%correct
1	4	4	4	1	17333	0.9	0.121	17.608	5.792	36.958	1.402	2.285	0.9	10.737	22.518	0.548	0.231	5	6406	36.958
1	3	4	5	1	17328	0.381	4.097	1.876	18.813	36.559	15.178	1.46	3.295	9.447	2.383	5.996	0.514	5	6335	36.559
1	4	4	5	1	17380	0.69	5.903	2.532	4.114	36.525	20.086	2.192	0.04	10.115	17.319	0.316	0.167	5	6348	36.525
1	2	4	5	1	17397	0.023	4.667	1.77	22.349	36.322	1.293	0.362	2.316	9.622	14.411	5.834	1.029	5	6319	36.322
1	1	5	5	1	17378	2.825	9.472	0.075	5.933	0.472	20.474	0	4.35	1.847	17.706	35.925	0.921	11	6243	35.925
1	2	4	6	1	17372	0.035	4.358	1.997	3.454	35.787	1.658	19.382	3.189	8.41	14.742	6.061	0.927	5	6217	35.787
1	4	4	10	1	17304	0.029	5.727	12.962	4.537	34.316	19.626	2.335	0.89	1.456	17.539	0.341	0.243	5	5938	34.316
1	4	4	2	1	17485	0.532	4.633	12.742	0.063	34.058	19.234	2.162	0.698	7.429	17.947	0.32	0.183	5	5955	34.058
...					...															
					99999985	8.333	8.334	8.333	8.333	8.333	8.334	8.333	8.334	8.335	8.333	8.333	8.333	2	27479970	27.480

Table: Sample of joint probability table for the OCCAM Fit of model *ACEHIZi* with p(margin) and p(rule) columns excluded as they were 0 for all rows.

Conclusion

More investigation and analysis needed, ideally with more mathematical training and/or the support of trained mathematicians.

References

Robert J. Lemke Oliver and Kannan Soundararajan. "Unexpected biases in the distribution of consecutive primes". In: *Proceedings of the National Academy of Sciences* 113.31 (July 2016), E4446–E4454. ISSN: 1091-6490. DOI: 10.1073/pnas.1605366113. URL: http://dx.doi.org/10.1073/pnas.1605366113.