STATISTICAL APPROACH TO REDUCTION OF STATE SPACE FOR HUMAN GENETIC DIVERSITY

ZULFIKAR MOINUDDIN AHMED

1. Refocus From 1000 Genomes Project

The work of the 1000 Genomes Project is great. Now I am looking at the snip allele frequencies for Chromosome 2 at the moment, and I will have the full Allele Frequency in a week or so.

It is very clear to me that independence is not a serious feature here, and I want a *vast* reduction of the dimension from N_{snips}. The method to do this is quite canonical since my Discrete Correlation algorithm will produce an $N_{snip} \times N_{snip}$ correlation matrix and I can use straightforward linear algebraic techniques to reduce dimension, counting eigenvalues, manipulating the correlation matrix this and that way. I won't go into specific techniques but anyone with experience with multivariate statistics knows a thousand things to do to reduce state space by producing uncorrelated subsets of dimensions. I don't know how much it will be but I expect $N_{blocksnips} << N_{snips}$ and the state space to become $2^{N_{blocksnips}}$. We can then be confident that we have this as the theoretically sound scientific model of possible genetic variety in human beings. Then we use Affy type fluorescence to find fast ways of measuring just the minimal subset of snips and then correlate these to every aspect of individual life on Earth including which exact passages from the Torah or Bible or Quran are actually useful for the person. This is not a frivolous issue. Every single aspect of a person's life can be correlated with core genetic uniqueness and we can fine-tune the life satisfaction of every person on Earth precisely in the future. Genetic uniqueness is the most powerful objective variable that human beings will ever possess and the resulting revolution in human affairs will literally turn half a century of technology revolution to purely sordid ghetto small affairs.

You think I am joking or exaggerating? Gautam Buddha could not appreciate how to handle the sufferings of my Beloved People the Human Race as much as genetic uniqueness and its applications can. Human Civilisation will have a new feature afterward. That feature will be Civilisation.

2. Gentle Testing Approach To Saturation of Binary Blocks for Snips

Human genetic variation is a very serious scientific problem, and since it is important, we want to get some understanding of the saturation of the 2^K binary spaces when K snip locations are randomly picked from the genome. So we want to pick $K = 10$ snip positions randomly, and generate a csv file with rows corresponding to either 0 for reference and 1 for the alternative allele. Then we can

Date: June 16, 2021.

check for the percentage of 2^K total combination that is saturated. If we do Monte Carlo with this algorithm eventually we ought to be able to have a good estimate of the percentage of saturation that we expect when $K >> 10$.

```
import vcf
import random

K = 10
min_idx = 0
max_idx = 100000
idx = []
for k in range(K):
    idx = idx.append( random.randint(min_idx,max_idx)

idx = sorted(idx)


def gv(sval):
    return(float(sval[-1]))

def push_vcf_reader( r, q):
    out = next(r)
    for w in range(q-1):
        out = next(r)
    return(out)

vcf_reader = vcf.Reader(filename='all.vcf.gz',encoding='utf-8')
samp_names = vcf_reader.samples

for sn in samp_names:
    row_data = ''
    for k in range(len(idx)):
        if k == 0:
            rec = push_vcf_reader(vcf_reader, idx[k])
        else:
            rec = push_vcf_reader(vcf_reader, idx[k]-idx[k-1])
        #print(str(rec.POS)+","+str(rec.REF) + "," + str(rec.ALT) + "," + str(rec.INFO['AF'][
        row_data = row_data + ',' + gv(rec.genotype(sn)['GT'])
    print(row_data)
```

Then we save to csv, pull up in R and do some work on R.

3. Astounding Lack of Perspective of Bill Gates

Bill Gates' comments about this are (a) he did not know whether Monte Carlo would converge, and this was 'trivia', and (b) he did not like the code organisation. I am just astounded by this. I have been doing quant work since 1995 from right after graduation from Princeton, and I have never actually heard of any senior person in any technical discipline at all not know about how Weak Law of Large Numbers allows Monte Carlo to produce estimates of functionals of unknown distributions

accurately. I just showed the world a *tractable path* to resolve an extremely difficult problem in biology, the estimation of actual state space for human diversity, using Monte Carlo methods that are tractable and can be assessed with computer resources that are not exorbitant, on a problem whose scientific importance is very high and whose resolution was unknown until just now, and this man Bill Gates calls this 'trivia' but does not consider issues of code organisation on research code to be 'trivia'. To say that I am totally flabbergasted is an exaggeration. How is this man considered even remotely educated in mathematics? Bill Gates ought to spend much more time on weak law of large numbers if he does not understand the principles of why Monte Carlo is routinely effective in Finance and Science. I will be quite honest: if this man interviewed with me for a top spot in technical matters I'd throw him out the window in 2 minutes.

4. Stepwise Progress

```
import vcf
import random

K = 10
min_idx = 0
max_idx = 100000
idx = []
for k in range(K):
    idx.append( random.randint(min_idx,max_idx))

idx = sorted(idx)
print('indices')
print(idx)

def gv(sval):
    return(float(sval[-1]))

def push_vcf_reader( r, q):
    out = next(r)
    for w in range(q-1):
        out = next(r)
    return(out)

vcf_reader = vcf.Reader(filename='all.vcf.gz',encoding='utf-8')
samp_names = vcf_reader.samples
records = [0] * len(idx)
for k in range(len(idx)):
    if k == 0:
        records[k] = push_vcf_reader(vcf_reader, idx[k])
    else:
        records[k] = push_vcf_reader(vcf_reader, idx[k]-idx[k-1])

print('obtained records')
for sn in samp_names:
```

```
    row_data = ''
    for k in range(len(idx)):
        #print(str(rec.POS)+","+str(rec.REF) + "," + str(rec.ALT) + "," + str(rec.INFO['AF'][
        rec = records[k]
        row_data = row_data + ',' + str(gv(rec.genotype(sn)['GT']))
        print(row_data)
    print(row_data)
```

Totally minor fixes, storing of the records first, etc. The code is still not fast enough to be usable, so I will have to massage it more.

5. Evaluation Code

```
binaryFraction<-function(X){
  tot <- 0
  nr <- dim(X)[1]
  done_rows<-rep(0,nr)
  for (j in 1:1024){
    check_val<-as.numeric(intToBits(j))[1:10]
    for (r in 1:nr){
      if (done_rows[r]){
        next
      }
      if (!(check_val[1] %in% unique(X[,1]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[2] %in% unique(X[,2]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[3] %in% unique(X[,3]))){
        done_rows[r]<-1
        next
      }

      if (!(check_val[4] %in% unique(X[,4]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[5] %in% unique(X[,5]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[6] %in% unique(X[,6]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[7] %in% unique(X[,7]))){
        done_rows[r]<-1
```

```
        next
      }
      if (!(check_val[8] %in% unique(X[,8]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[9] %in% unique(X[,9]))){
        done_rows[r]<-1
        next
      }
      if (!(check_val[10] %in% unique(X[,10]))){
        done_rows[r]<-1
        next
      }
      if (norm(check_val-as.numeric(X[r,]),type="2")<0.01){
        tot<-tot+1
        done_rows[r]<-1
        break
      }
    }
  }
  tot/1024
}
> source('~/thy/papers/bincov.R')
> binaryFraction(X)
[1] 0.002929688
```

This is just one random Monte Carlo run. But this is most likely indicative. The allele state space is 0.3% here and most likely going to be in this range for $K = 10$.

6. Initial Estimate of Diversity Density

We expect initially around 3^{10} of binary code of length 100 to be filled up for 100 snip locations. That's $\exp(-58.32)$ of 2^{100}.

7. This Estimate Is Very Close To Tight

```
> exp(-58.33+100*log(2))
[1] 58966.1
> 59000*60000
[1] 3.54e+09
```

This sort of estimate is getting close to the living population of the human race, which is very tight. It's not right yet, but this is what we're looking for in human genetic diversity, some sense of realistic values in 10-50 billion people.

8. Rough Estimations

I will be doing more careful Monte Carlo simulations later on. Let me tell you how I think of estimating these things.

I expect the density for 1024 to be 0-10 say, not more. Let's say it is $0 < d < 10$. Then I work as follows. Per 100 snips I get

$$y(d) = 10\log(d) - 100\log(2)$$

This will be something like $y = -52$. Then I get

$$T(d) = exp(y(d)) \times 2^{100} \times 60000$$

as the estimates of the total density. Let's get some representative values.

d	T(d) in billions
1	0.00
2	0.06
3	3.54
4	62.91
5	585.94
6	3627.97
7	16948.51
8	64424.51
9	209207.06
10	600000.00
11	1556245.48
12	3715041.85
13	8271509.51
14	17355279.30
15	34599023.44

9. Another Estimate

I get initial values for saturation as

$$[0.0039, 0.0166, 0.00097]$$

The mean translates to 7.33/1024. This then gives a total estimate of 27 trillion possible people. We'll get finer results later but this looks reasonable. And this is extremely important for Science for Human Genetic Diversity. This is much more reasonable than 2^{600000} possibilities. Now 27 trillion is reasonable, and that is the problem with just using allele probabilities. This is much better estimate, total 27 trillion possibilities. With many more Monte Carlo runs we will have more reliable estimates.

10. Improved Estimates

```
> vs<-c(0.0039,0.0166,0.00097,0.0009765,0.01074,0.000977,0.00488,0.0097656)
> (mean(vs)*1024)^10*60000/1e9
[1] 5435.743
```

This estimate is 5.4 trillion possible human beings for possible genetic diversity. Monte Carlo takes time, so the runs are few so the estimate moves around. Regardless I think 5.4 trillion is a roughly correct value. We do not expect now to see 1000 trillion. In a later section I consider the implication for our *moral nature*. Genetic diversity obviously affects all aspects of our lives as Human Race. I have been looking at regularity of moral nature of Man across the globe. We can make the

qualitative judgment here and expect that our moral nature will reach equilibrium in the future centuries and also that immoral people will simply die out.

11. Charles Murray's p. 203 quote

I like Charles Murray's book. And it's reasonable till you come across Bill Gates. He has a straw man orthodoxy view on p.202.

"The system is rigged in favour of heterosexual white males. The privilege is accorded them accounts for who gets around and who is kept at the bottom." (Charles Murray, *Human Diversity*)

Now I am a Princeton graduate, and I don't much care about what the heterosexual whites are up to. I have my own interests. I did not actually adhere to the orthodoxy presented here. I don't know if this orthodoxy is true or not. What I do know is that Bill Gates believes in this strongly, is totally outrageously racial, and is explicitly continuously plotting to harm non-white people and keep them down. He won't succeed, because I am superior to him and I'll wipe the fuck out of existence sooner or later, but what Charles Murray calls 'orthodox opinion' is not orthodox opinion but Bill Gates' explicit agenda for decades.

12. Schroedinger Was A Scientist of First Rank

Erwin Schroedinger was an extraordinary scientific genius of the first rank. Between Einstein and him, he was the better *scientist* but Einstein was the deeper philosopher of science, and Einstein's *taste* was better. Yes, it matters. They were both actually good. I overthrew them it's true but without much respect lost for either. You see, in a sense Four-Sphere Theory was a cheap shot, because I had measured cosmological constant right there, and I could just follow the mathematical path and go through my 'and therefore too bad for Einstein and also too bad for Schroedinger' and so on. Neither Einstein nor Schroedinger had this sort of luxury where one has all the pieces arranged just so for me and I just put them together and voila, Nature in her glorious dressing smiling at the foolishness of Man. My difficulties are not as much scientific as much as sociological, such as racial prejudices that were rankled when rather scientifically untalented people like Bill Gates were unable to resist giving up their lives and sacrificing them by irritating me by trying to kill me. As far as I know, neither Einstein nor Schroedinger had any rich little fuck with strong racial views try to actually do away with them with great zeal.

13. Human Moral Nature Is Tightly Constrained For Our Future

I was the first to discover strong regularities across the globe of our moral values that are affected by ethnicity by no more than 6.5-9.5%. Here we are seeing that the potential genetic diversity altogether for future humans is also quite constrained. Let's consider an estimate from around 2-100 trillion. That's not a lot of *space* to produce vastly exotic moral machinery in our genetic make-up. Today we have some moral degenerates, evil psychopaths like Bill Gates, but we will not be able to transform ourselves into a *race* of parasites like Bill Gates even if that were our fondest wish. It so happens that I have also showed that Aristotle was right and we have higher life satisfaction when we are virtuous. So I want to make the inference that there is an *equilibrium moral nature* that we have not reached only because of the primitive manner in which we have been ignorant of the universality

of moral natures that had existed for centuries hidden and not measured before World Values Survey and other recent efforts.