Compressed Sensing Application to Droplet Digital PCR (Notes)

Cassidy Mentus

September 5, 2020

1 Introduction

1.1 Compressed sensing

The basic setting of "compressed sensing" is that you have a signal (denoted x) that you wish to accurately measure using a small number of measurements (denoted y). Surprisingly, it turns out that this is achievable a rate far below that would be predicted by theory if we make an extra assumption: the signal has a sparse representation as 'atoms'- i.e. most of the atoms are zero. For testing the presence of COVID-19 the 'atoms' are samples. The non-zero (positive) atoms are COVID-19 positive, zero atoms are COVID-19 negative and the number of non-zero atoms (the sparsity) is the prevalence. When the prevalence is low you can use a "sensing matrix" (denoted Phi) that tells you how much of each atom to pool into a measurement. Given appropriate assumptions on the noise in measuring x we are guaranteed to recover the true signal and, furthermore, the set of positive atoms with certain accuracy.

1.2 Biological/medical application

- Droplet Digital PCR divides one sample into thousands of microdroplets.
- More than one probe with different or the same colored dye can be used to detect nucleotide sequences.
- Group testing has been used in the past to efficiently identify disease carriers when prevalence is low.
- Compressed sensing is connected to group testing because sparse expanders are useful for both. Also we can use boolean group testing in the form of COMP to narrow down the support.
- This can be applied to people, but it can be applied to other things we may want to assay such as sewage or other biomes.
- This work is useful for other micro-arrays and technology that will come out in the future.

1.3 What is achieved by this paper

- Introduce compressed sensing with Poisson noise to ddPCR based assay. This is also a useful foundation for qPCR because it is a good way to model noise from dilution.
- Introduce a new (and best performing) penalty function, the differential entropy of a poisson distribution.
- Compare ℓ_1 reg, $\ell_{\frac{1}{2}}$ reg, H- reg performance.
- Apply sparse left-expander matrices derived from deterministic codes used in Tapestry (Ghosh et al) and PBEST (Shental et al).
- Optimize non-uniform sensing matrices to have low mutual coherence. Demonstrate their effectiveness although they are outperformed by the expander matrices.
- Introduce Tensor (Kroneckerized) Compressed Sensing to pool to detect a plethora of rare DNA sequences in a populations.

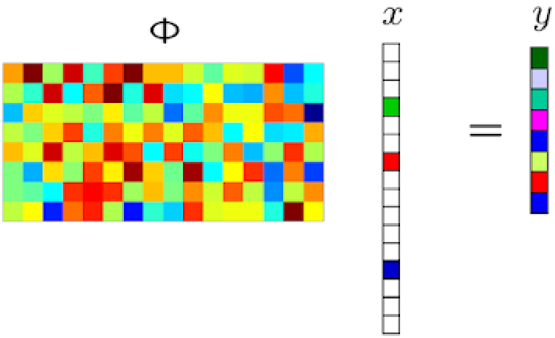

Figure 1.1: Image credit: http://informationtransfereconomics.blogspot.com/2017/10/compressed-sensing-and-information.html

To do

- Study how different choice of a and τ improves our estimation along the lines of viral concentration, accuracy, False Positives False Negatives.
- What percentage of the false negatives are "unavoidable" caused by dilution–not by the algorithm.
- qPCR: Consider the poisson variable as a hidden variable. Then the addition of "multiplicative noise" will give the observed variable. Develop a MLE estimate to solve this problem. Maybe as a log normal?
- Optimize the penalty hyperparameter.

2 Compressed sensing with Poisson noise model

1. Sparse signal recovery
2. Objective function Poisson log-likelihood with penalty $a(x)$ and scalar τ: $\sum_{i=1}^{M} \left[[\Phi x]_i - y_i \log [\Phi x]_i\right] + \tau a(x)$

2.1 Penalty functions

3 Sensing matrix design

Valid matrices have low mutual coherence and are so that $\sum x_i \leq \sum_i [\Phi x]_i$.

1. Kirkman triple (Tapestry Ghosh et al) ($N = 238$, $M = 42$)
   (a) Sparse expander matrix. Designed so that each sample is pipetted into only 3 pools. Therefore lab-personnel can use it.
   (b) $\mu = \frac{1}{3}$ $\bar{\mu} = .07$ where μ is the mutual coherence (the maximum of the normalized dot-products) and $\bar{\mu}$ is the average of the dot products of unit-vectors.
2. PBEST Solomon-Reed code ($N = 384$ $M = 48$)
   (a) Sparse expander. Each sample goes into 6 pools so it is best to use a liquid handler
   (b) $\mu = \frac{1}{3}$ $\bar{\mu} = .1227$
3. μ-optimized sensing matrix
   (a) For comparison, we apply a non-convex μ optimization algorithm to get a locally optimal matrix.

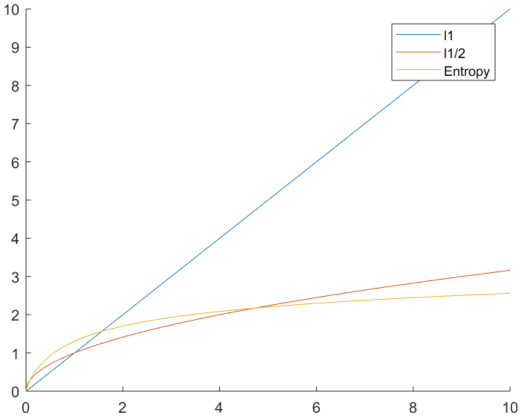

Figure 4.1: Plots of the penalty functions. The differential entropy of a poisson random variable with mean x is plotted. The entropy function grows at a much slower rate (logarithmically) causing it to have almost 0 shrinking effect at any scaling.

(b) These are still sparse matrices.

(c) $\mu = .39$ $\bar{\mu} = .19$

(d) Normalize by maximum row sum- so there is extra dilution, but not that much. In the future it will be best to get a rotation matrix to make their mean direction close to $\mathbf{1}_M$ (direction of the all 1s matrix)

4 Penalty functions

The types of penalty function are ordered in increasing order of "sparsity" and decreasing order for reconstruction accuracy.

4.1 ℓ_1 recovery convex problem $a(x) = \|x\|_1$ (3rd best option)

- Convex optimization converges to global optima.
- Shrinks the estimated viral concentrations in a counter intuitive manner because Total viral concentration $\propto \sum_i [\Phi x]_i$.
- Can estimate the support but a sparser penalty will be better $\|x\|_{\frac{1}{2}}$ or $H(x)$.

4.2 $\ell_{\frac{1}{2}}$recovery non-convex problem $a(x) = \|x\|_{\frac{1}{2}}$ (2nd best option)

- This is a sparser penalty because it has a sharp cusp at zero and its slower growth than the ℓ_1 norm.
- Higher τ penalty scalar improves false positive rate but it shrinks the values although it is
- Has a good reason. Solving for the MLE finds many low viral load cases that model the fluctuations between high viral load pools. Therefore the distribution in pools given $\hat{x}$ (the estimate) are not indepent. Low concentration gets coupled to high concentration.

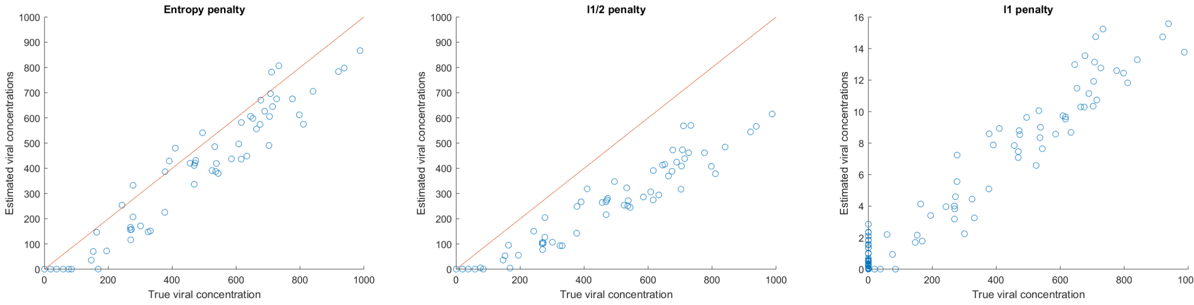

Figure 6.1: The entropy does not cause any shrinking while accurately finding the sparse basis. The $\ell 1/2$ penalty also accurately finds the basis, but it shrinks the predicted concentrations. Also there is more noise. The ℓ_1, for some reason (?) performs much worse. Maybe the optimization algorithm has to be fixed. If we consider this as a MAP optimization, then the ℓ_1 assumes a Laplace prior, while the entropy has approximately a power law. These fit the assumptions of raginski's paper. From left to right we use $\tau = 30, 5, 10$.

4.3 Entropy penalty $\sum_{i=1}^{N} H(x_i)$ (1st best option)

According to [1] the entropy of a Poisson distribution with mean x_i can be expressed

$$H(x) \quad = \quad -x\log(x/e) + \int_0^1 \frac{1 - e^{-xz} - xz}{z\log(1-z)}dz.$$

The derivative of this expression with respect to x is

$$\frac{d}{dx}H(x) \quad = \quad -\log{(x)} + \int_0^1 \frac{ze^{-xz} - z}{z\log(1-z)}dz.$$

This is a sparse penalty function and we can optimize. It is also connected to minimum description length. We expect this to cause less shrinking as $H(x) \approx \frac{1}{2}\log(2\pi e x)$ for large x.

5 Optimization algorithm

The best options are iteratively re-weighted least squares or sparse bayesian learning. We find that iteratively reweighted least $\|\cdot\|_1$ has good performance and it is very easy to implement. We repeatedly define hyperplanes supported by the negative log-likelihood at the current estimate as an upper bound. The CVX matlab package is then used to optimize each upperbound function.

6 Preliminary results

6.1 Kirkman triple $N = 238$ $M = 42$

We sample 10 populations with 238 samples and 6 positive cases. The viral concentrations are independent uniform $[0, 1000]$ random variables.

6.2 μ-optimized $N = 238$ $M = 42$

7 Many samples, many rare nucleotide sequences: Tensor Compressed sensing.

Consider the Kirkman triple bipartite graph with the left vertex set (samples) consisting of 153 nodes and the right vertex set containing 33 nodes (tests). Each pool has 17 samples in it. We are testing 153 samples

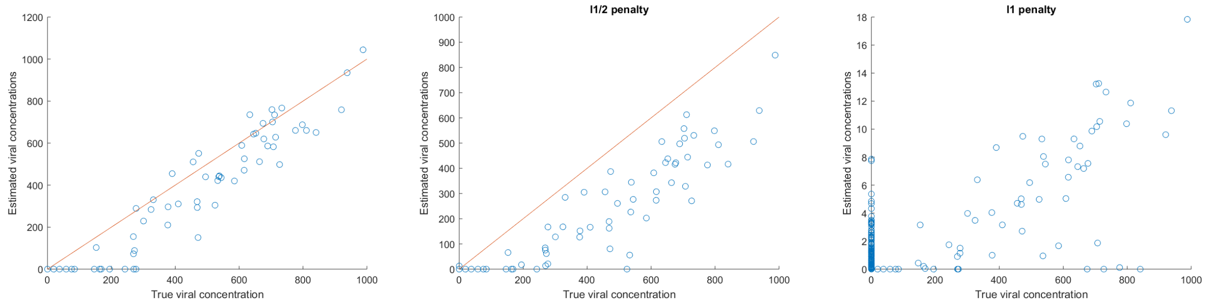

Figure 6.2: The estimated viral concentrations using a μ-optimized matrix with penalties (from left to right) entropy $\ell_{1/2}$ and ℓ_1. Even though the mutual coherence of the matrix is not much higher than that of the Kirkman triple matrix, the results are noisier and have more false negatives. Still, the results are good for the entropy penalty which we note, estimates viral concentrations with the lowest error. We used the same random populations and viral concentrations as the Kirkman experiment

for one DNA sequence. We can reverse the role of samples and DNA sequences and instead test one sample to see if it has contains some of the 153 different low-prevalence DNA sequences using 33 tests. This may have applications in biological research and for diagnostics if state of the art subdivided assays improve in speed and capacity.

We can extend this idea to test N_{samples} sample for N_{probe} DNA sequences. This is done through **tensor compressed sensing** using an expander graph for multiplexing probes $\Psi \in \mathbb{R}^{M_{\text{samples}} \times N_{\text{samples}}}$ and an expander graph for multiplexing samples $\Phi \in \mathbb{R}^{M_{\text{probes}} \times N_{\text{probes}}}$. An important nuance about Ψ is that it will not be normalized because we are assuming that adding probes does not change the concentration. **We must find out if this will invalidate the reconstruction error given by Raginsky et al.** Let x_{ij} denote the concentration of DNA containing sequence j in sample i and denote the $N_{\text{sample}} \times N_{\text{probe}}$ matrix as x. Then pooling concentrations will be $\Phi x \Psi^T$. This left and right multiplcation by two matrices is a linear function of x so we can re-write this as a matrix times a vector and use the methods above. This is equivalent to forming a sub-graph of the tensor graph where the left nodes are paired with the left nodes and the right nodes are paired with the right nodes. This new graph is also a good expander.

7.1 To-do:

- Rigorously state and either show or find out if correctly stating that this tensor of expanders is a good left expander.
- Write code for a small Kirkman triple on the order of 100 variables pooled in 20-50 pools and take $\Phi = \Psi$. That is, use the same matrix to pool samples as well as DNA probes.

References

[1] Mahdi Cheraghchi. Expressions for the entropy of binomial-type distributions. In *2018 IEEE International Symposium on Information Theory (ISIT)*, pages 2520–2524. IEEE, 2018.