FLORIDA ATLANTIC UNIVERSITY

DATA SCIENCE, ANALYTICS, AND ARTIFICIAL INTELLIGENCE CONFERENCE

SATURDAY
NOVEMBER 14, 2020

FAU.EDU/DATA

BREAKOUT SESSION

PRESENTATION DESCRIPTION

CHARLES E. SCHMIDT COLLEGE OF SCIENCE

fau.edu/data

Group Testing and Compressed Sensing for COVID-19 Using ddPCR

Cassidy Mentus PhD UCLA 2019 (Florida resident)

Note to committee:

This is rough and some of the citations are lost because I had to use word since I am sending this off for some funding, so please be kind when judging it. Thank you very much again for the opportunity to present at your conference. I hope that you consider this a nice practical twist on COVID-19 related research.

B. Background

B.1. Two-stage group testing methods

To ramp up to the two-stage methods that we apply using the Group-Well app we explain the **basic group testing method.** Let G be a fixed group size. For example $G = 4$ as is currently approved for qPCR COVID-19 assays. The method has two steps with the tests in the second step determined by stage 1 results as follows:

1. **Screening stage:** Divide the samples into groups of size G. Assay the pooled samples for each group.
    a. All samples in negative groups are screened as negative.
    b. For a positive group, any person may have COVID-19. Therefore we say they are **potentially positive**.
2. **Individual testing**: Test the potentially positive samples one by one.

The set of potentially positive individuals can be quickly narrowed down in one round if we allow groups to overlap. We call this collection of strategies **two stage group-testing.** To define a two stage testing strategy on a population of size N we specify, M the number of pools, L the number of pools each sample will go into and the group size G. The basic group testing method example above for pooling 4 at a time is described as pooling where each sample is only included in $L = 1$ pools, the number of pools $M = \frac{N}{4}$ and $G = 4$.

Each sample will go into the same number of pools and the group sizes will be equal. This is important because we can estimate the drop in sensitivity caused by dilution for each viral concentration. An important aspect of our pools is that the intersection of two pools contains at most 1 sample. This property reduces the number of potential positives. The pooling schemes will be favorable for both total number of tests used if an extra round is used for confirmatory testing and for accurate decoding using compressed sensing described in the next section.

B.2. 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 at 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 Φ) that tells you how much of each sample to pool into a measurement (pooled assay). Given appropriate assumptions on the noise in measuring x we are guaranteed to recover the true signal and, furthermore, the set of positive samples with certain accuracy.

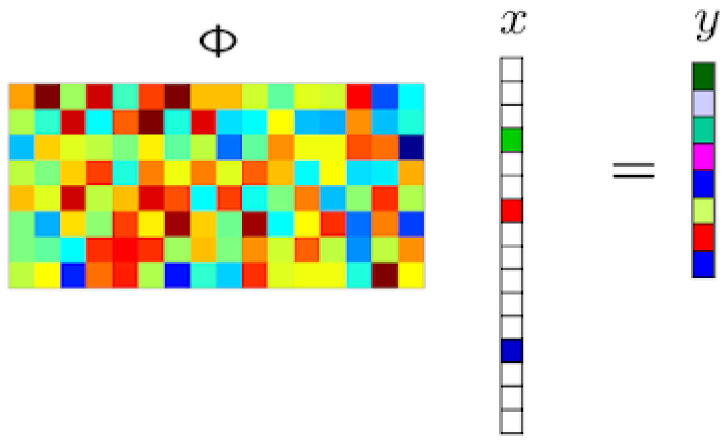
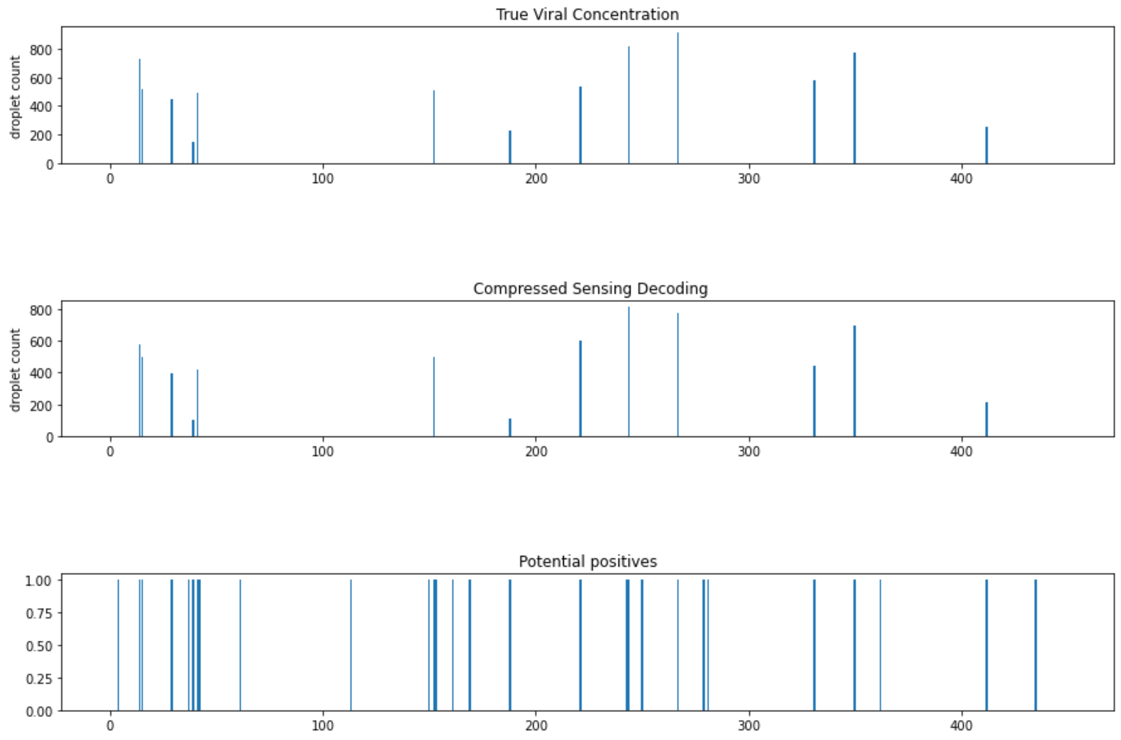


Image credit: http://informationtransfereconomics.blogspot.com/2017/10/compressed-sensing-and-information.html

For pooling PCR assays, the matrix Φ is constructed normalizing the rows of the 0-1 entries representing whether or not a sample is in a pool. We can model the pool concentrations Poisson random variables. For this exact setting and our pooling designs there are theoretical bounds on our ability to recover the true signal https://arxiv.org/pdf/1007.2377.pdf . The next figure shows promising a preliminary application of compressed sensing on synthetic viral concentrations and Poisson r.v. pool positive droplet count.

We simulate 450 samples containing 13 positive samples with viral loads chosen uniformly over [0,1000] (top). After one round of testing 90 pools containing 15 samples each, by declaring all samples in negative pools as negative, we narrow down the potential positive cases to 25 (bottom). We apply a compressed sensing with Poisson noise on the pooled measurements to decode individual samples’ viral concentrations (middle). For this example, we achieve perfect recovery of the set of COVID-19 + samples with close approximations to their viral concentrations.



Bar-plots for true viral concentrations (top), compressed sensing based decoding (middle) and the potential positives after one round of testing. There are 450 samples and we assume that 13 are COVID-19 (+). Decoding by minimizing the penalized negative log-likelihood (VII.B.2.3) recovers not only the set of positive cases and closely approximates the samples viral concentrations.

B.2.1 Mathematical framework

The mathematical setting of our application of compressed sensing consists of signal vectors $x \in (R_{\geq 0})^N$, measurement vectors $y \in (Z_{\geq 0})^M$ and a sensing matrix $\Phi \in R^{M \times N}$. In our application, x represents the true viral concentrations of each sample, y is droplet count by ddPCR and Φ is the pooling matrix with each row normalized to sum to 1. That is,

$$\Phi_{ij} = \frac{1}{G} \ if \ j \in \text{Pool } i$$

and

$$\Phi_{ij} = 0 \text{otherwise.}$$

We assume that the number of droplets y_i in pool i is a poisson random variable at lower concentrations with mean the average concentration in each pool. This is expressed as the multivariable probability distribution

$$y \sim Poisson(\Phi x).$$

At higher concentrations it is more accurate to model droplet counts with a normal or binomial distribution, and we will research these different assumptions.

In our studies we will investigate compressed sensing using the sum of least squares and a probabilistic formulation that takes into account the poisson noise.

B.2.2 Compressed sensing using the sum of least squares

This is the most common description of compressed sensing. WeUsing the sum of least squares least squares, compressed sensing in a testing scenario assumes that the prevalence of COVID-19 (# COVID-19 positive samples) in the test population is $\leq s$ for some small s. If the sensing matrix satisfies a restricted isometry property or has low coherence, then, depending on the model for noise solving the convex optimization problem

$$\hat{x} = arg \min_{x \in (R_{\geq 0})^N} \| \Phi x - y \|_2^2 + \tau \| x \|_1$$

is guaranteed to closely approximate both the true viral concentrations x^* and the set of COVID-19 + samples. That is, $\{i: \hat{x}_i > 0\}$ is very close to the true set of COVID-19 + samples. This minimization is usually referred to as LASSO.

Our pooling matrices have an ℓ_1 restricted isometry property (RIP-1)

$$(1 - \epsilon) \| x \|_1 \leq \| \Phi x \|_1 \leq \| x \|_1$$

for some small ϵ on the set of all vectors x that are sparse enough (have low prevalence). Then the sensing matrix has low coherence (the maximum of the dot product over all pairs of distinct rows) $max_{1 \leq i < j \leq N} \phi_i \cdot \phi_j \leq \frac{1}{G^2}$ as a result of any two pools' intersection containing at most one sample.

B.2.3 Compressed sensing with Poisson noise

Although compressed sensing solving the LASSO optimization problem has been useful for decoding pools using qPCR , there has yet to be any research applying compressed sensing to pooling with ddPCR. For pooled measurements with < 1000 droplets (5% of the maximum number of droplets) in an ideal model, the measurements are Poisson random variables. Our pooling matrices fit the assumptions Raginsky et al guaranteeing accuracy of approximations minimizing the penalized log-likelihood function

$$\hat{x} = arg \min_{x \in (R_{\geq 0})^N} \sum_{i=1}^{M} \quad y_i - y_i log[\Phi x]_i + \tau a(x)$$

for $\tau > 0$ where $x \in [\Phi x]_i$ represents the i^{th} entry of the concentrations of pools calculated assuming sample concentrations x and $a(x)$ is a penalty function. The penalty function can be chosen to penalize sets of viral concentrations with higher prevalence. One option is to use the sparse regularization term $a(x) = \| x \|_1$ but for equally sized groups, $a(x^*) \approx \frac{G}{L} \sum_{i=1}^{M} \quad y_i$. So shrinking the overall sum leads to less accurate viral concentrations although, in our experience, decodes the set of COVID-19 + accurately. An even better choice we feel is

$$a(x) \ = \| x \|_{\frac{1}{2}} := \sum_{i=1}^{N} \quad \sqrt{x}.$$

We can think of this as modelling the viral concentrations (in droplets) as poisson random variables with a single standard deviation parameter. This is then the ℓ_1 norm of the standard-deviations vector. We find that it may prevent over-fitting for pools with high viral concentrations because the unpenalized MLE often lead to small positive terms on order of the noise. Penalizing the sum of “noises” encourages noise to be concentrated into higher viral-load cases in the decoded viral concentrations.

B.3 Related work

There have been several applications of pooling and compressed sensing to COVID-19 testing, but they have only been carried out on qPCR tests of samples or numerical simulations. To our knowledge, we will be the first to apply compressed sensing to ddPCR. In [cite] least squared error is minimized to denoise the qPCR pooled assays. They use deterministic pooling matrices derived from Kirkman triples. We will include these pooling matrices in our studies due to their success on qPCR samples. In pooling matrices derived from Reed-Solomon and random Bernoulli matrices are used. Both use liquid handlers demonstrating their efficacy for pooling with qPCR. We wish to repeat this state of the art technique. The connection between compressed sensing and pooling for DNA sequence detection has been discussed since at least 2009 . Still ddPCR has yet to be augmented using compressed sensing techniques.