Fast and Efficient Data Science Techniques for COVID-19 Group Testing

Varlam Kutateladze[1] Ekaterina Seregina[2]

[1,2]University of California Riverside

OCLB ASA 2020 Q3 Meeting

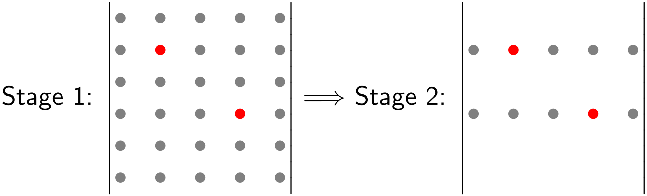
Group Testing

The New York Times

Five People. One Test. This Is How You Get There.

Nebraska is testing more people with the tests it has. The technique is simple.

Dorfman 1943:



Stage 1: $\Longrightarrow$ Stage 2:

30 individuals, 2 infected, $6 + 10$ tests.

- Sterrett 1957, Sobel et al. 1959, many more

Motivation

> *"Because samples are pooled together, ultimately fewer tests are run overall, meaning fewer testing supplies are used, and results can be returned to patients more quickly in most cases."*
>
> **FDA**

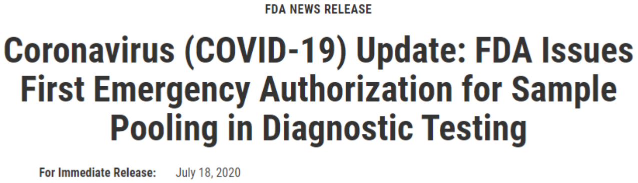
Why do group testing?

- Increased testing throughput
- Limited use of chemical reagents
- Higher overall testing capacity

Biomedical considerations:

- Dilution not too severe (Hogan et al. 2020, Yelin et al. 2020, Abdalhamid et al. 2020, Mutesa et al. 2020)
- Successfully used for HIV (Emmanuel et al. 1988), influenza (Van et al. 2012), malaria (Taylor et al. 2010), etc.

FDA Emergency Use Authorization



FDA NEWS RELEASE

Coronavirus (COVID-19) Update: FDA Issues First Emergency Authorization for Sample Pooling in Diagnostic Testing

For Immediate Release: July 18, 2020

> *"This EUA for sample pooling is an important step forward in getting more COVID-19 tests to more Americans more quickly while preserving testing supplies. Sample pooling becomes especially important as infection rates decline and we begin testing larger portions of the population."*
>
> **FDA Commissioner Stephen M. Hahn, M.D.**

- Pooling test performance should have ≥85% percent positive agreement (PPA) when compared with the same test performed on individual samples.

Adaptive vs Non-adaptive

Adaptive:

- ✓ Multiple stages
- ✓ Non-overlapping groups
- ✓ Testing procedure depends on previous test results

Non-adaptive:

- ✓ Single stage
- ✓ Overlapping groups
- ✓ Testing procedure does not depend on previous test results

This study

- Non-adaptive techniques
- Propose a simple method based on ℓ_1-norm sparse recovery

RT-qPCR

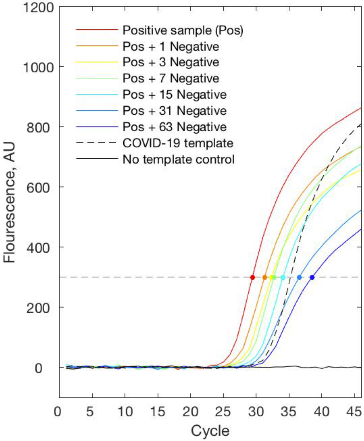
Reverse transcription quantitative polymerase chain reaction (RT-qPCR).

- Target cDNA is amplified exponentially for up to ~ 40 cycles.
- If fluorescent signal crosses a threshold before a certain number of cycles, the patient is declared positive.
- **Output**: Cycle threshold (CT), i.e. cycles completed before crossing the threshold.

Many algorithms do not take the quantitative information into account!

`https://tinyurl.com/y5se6w4n`

RT-qPCR



Source: Yelin et al. 2020

Problem Formulation

We have n individuals, k are positive. Want to identify with $m \ll n$ tests.

How to pool? Design an $m \times n$ matrix A.

$$\mathrm{A} = \begin{bmatrix} \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & a_{ij} & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \end{bmatrix} \qquad \mathrm{x} = \begin{bmatrix} \cdot \\ \cdot \\ x_j \\ \cdot \\ \cdot \end{bmatrix}$$

a_{ij} $= 1$ if individual j included in group i, $= 0$ otherwise
x_j $= 1$ if individual j positive, $= 0$ otherwise

✓ x could also be RT-qPCR quantitative readouts!

We observe $\boxed{\mathrm{y} = g(\mathrm{A}, \mathrm{x}, \epsilon) = \mathrm{A}\mathrm{x} + \epsilon}$, want to infer x.

Pooling Matrix Design

How to design A? Constant column weight design (Aldridge et al. 2016).

$$\mathrm{A} = \begin{bmatrix} 1 & 0 & 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 1 & 0 & 1 \\ 1 & 1 & 0 & 0 & 1 & 0 \\ 1 & 0 & 0 & 1 & 1 & 0 \\ 0 & 0 & 1 & 0 & 1 & 1 \end{bmatrix}$$

Columns of A have up to L ones, randomly filled by bootstrapping.

- ✓ Avoid too much dilution
- ✓ Better performance
- ✓ Theoretical justification

ℓ_1 sparse recovery

How to infer x? Want to solve:

$$\min_{\mathrm{x}\in\mathbb{R}^n} \quad \|\mathrm{x}\|_0 \quad \text{s.t.} \quad \|\mathrm{Ax}-\mathrm{y}\|_2 \leq \epsilon,$$

Equivalent to Basis Pursuit Denoising if A is RIP:

$$\min_{\mathrm{x}\in\mathbb{R}^n} \quad \|\mathrm{x}\|_1 \quad \text{s.t.} \quad \|\mathrm{Ax}-\mathrm{y}\|_2 \leq \epsilon,$$

Lasso:

$$\boxed{\min_{\mathrm{x}\in\mathbb{R}^n} \quad \|\mathrm{Ax}-\mathrm{y}\|_2^2 + \lambda \|\mathrm{x}\|_1}$$

Add $\mathrm{x} \geq 0$ constraint.

RIP

Definition

An $m \times n$ matrix A satisfies k-Restricted Isometry Property if $\exists \delta_k \in (0, 1)$:

$$(1 - \delta_k) \|\mathrm{x}\|_2^2 \leq \|\mathrm{A}\mathrm{x}\|_2^2 \leq (1 + \delta_k) \|\mathrm{x}\|_2^2 ,$$

for all k-sparse $\mathrm{x} \in \mathbb{R}^n$ (Candes et al. 2006, Donoho 2006).

Lemma

An $m \times n$ matrix A with constant column weight design satisfies RIP for some integer $L > 0$.

Advantages

Some benefits of this approach:

- One-round
- $m = O(k \log(n))$
- Inputs real-numbered readouts
- Reconstructs viral loads
- Works well with noise

Other non-adaptive algorithms:

- COMP (Combinatorial Orthogonal Matching Pursuit)
- DD (Definite Defectives)
- CBP (Combinatorial Basis Pursuit)
- SCOMP (Sequential COMP)

Comparison

Negative/Positive identification

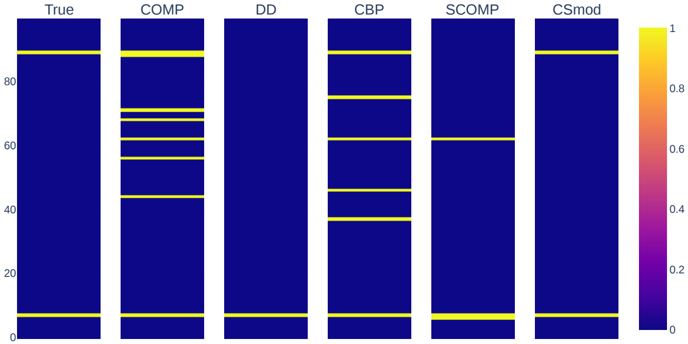


Figure 1: $n = 100$, $k = 2$, $m = 20$

RMSEs

$$\text{RMSE} = \frac{\|\mathrm{x}-\hat{\mathrm{x}}\|_2}{\|\hat{\mathrm{x}}\|_2}$$

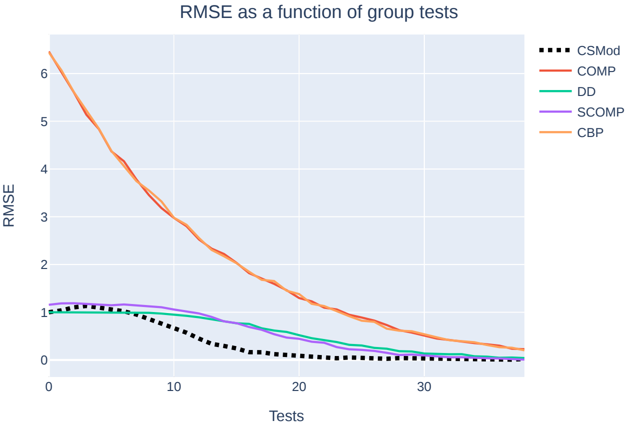


Figure 2: $n = 100$, $k = 2$, 1000 Monte Carlos

Sensitivity

Sensitivity = ratio of identified positives to all true positives

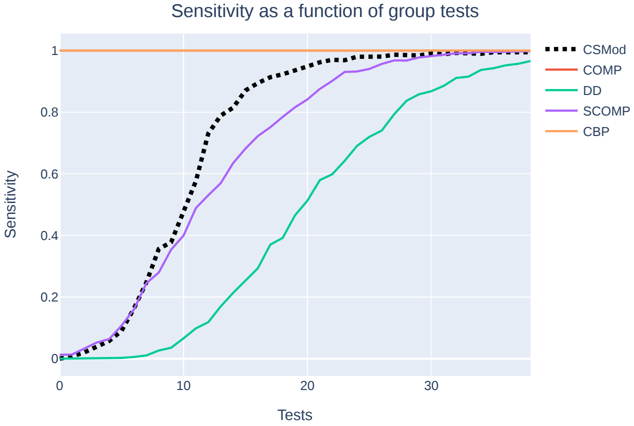


Figure 3: $n = 100$, $k = 2$, 1000 Monte Carlos

Specificity

Specificity = ratio of identified negatives to all true negatives

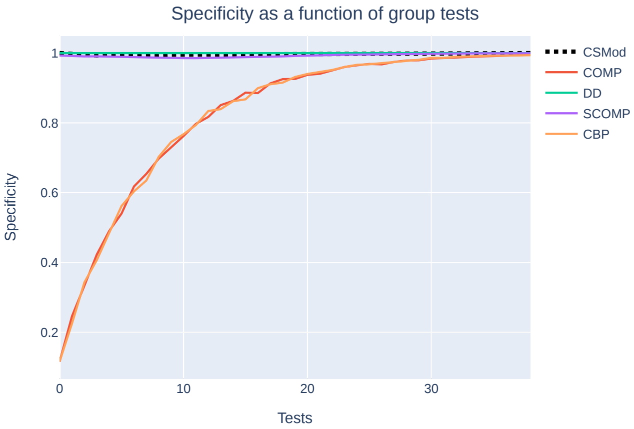


Figure 4: $n = 100$, $k = 2$, 1000 Monte Carlos

ROC curve

ROC curve for CSMod, thresholding Lasso estimates.

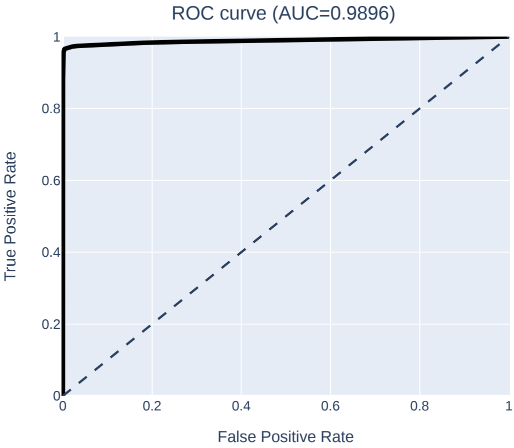


Figure 5: $n = 100$, $k = 2$, 1000 Monte Carlos

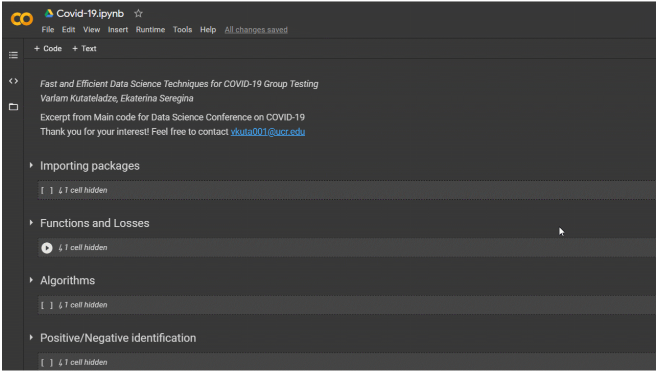
Improvement factor

Improvement factor $= \frac{n}{\mathbb{E}(\#\text{ of tests})}$ for 95% specificity & sensitivity.

	$\frac{k}{n} = 2\%$	$\frac{k}{n} = 4\%$	$\frac{k}{n} = 6\%$
Individual	1.00	1.00	1.00
Dorfman	3.37	2.60	2.15
COMP	4.53	2.80	1.96
DD	2.80	1.99	1.49
CBP	4.60	2.81	1.93
SCOMP	3.81	2.48	1.78
CSMod	5.11	4.01	3.42

Table 1: Improvement factors for three different prevalence rates, averaged over 1000 Monte Carlos

Google Colab link



Fast and Efficient Data Science Techniques for COVID-19 Group Testing
Varlam Kutateladze, Ekaterina Seregina

Excerpt from Main code for Data Science Conference on COVID-19
Thank you for your interest! Feel free to contact vkuta001@ucr.edu

Importing packages

[ ] ↳ 1 cell hidden

Functions and Losses

↳ 1 cell hidden

Algorithms

[ ] ↳ 1 cell hidden

Positive/Negative identification

[ ] ↳ 1 cell hidden

`https://tinyurl.com/y4vo86sb`

Similar approaches

See Yi et al. 2020 and Ghosh et al. 2020.

Key differences:

- Pooling matrix addresses current challenges and is flexible in size, shown to be RIP whp
- Different noise model
- Additional constraints: nonnegativity, less dilution
- Comparison with other algorithms

Main Takeaways

- Group testing could be beneficial at low disease prevalence rates
- ℓ_1 recovery works and is theoretically justified
- Fast and efficient, $m = O(k \log(n))$

Good resources:

- Chris Bilder website: `http://chrisbilder.com/grouptesting/`
- Book: Du et al. 1999
- References

Thank you! Contact: varlam@kutateladze.com

References I

Abdalhamid, Baha et al. (Apr. 2020). “Assessment of Specimen Pooling to Conserve SARS CoV-2 Testing Resources”. In: *American Journal of Clinical Pathology* 153.6, pp. 715–718. ISSN: 0002-9173. DOI: 10.1093/ajcp/aqaa064. URL: https://doi.org/10.1093/ajcp/aqaa064 (cit. on p. 3).

Aldridge, M., O. Johnson, and J. Scarlett (2016). “Improved group testing rates with constant column weight designs”. In: *2016 IEEE International Symposium on Information Theory (ISIT)*, pp. 1381–1385 (cit. on p. 9).

Candes, E. J., J. Romberg, and T. Tao (2006). “Robust uncertainty principles: exact signal reconstruction from highly incomplete frequency information”. In: *IEEE Transactions on Information Theory* 52.2, pp. 489–509 (cit. on p. 11).

Donoho, D. L. (2006). “Compressed sensing”. In: *IEEE Transactions on Information Theory* 52.4, pp. 1289–1306 (cit. on p. 11).

Dorfman, Robert (Dec. 1943). “The Detection of Defective Members of Large Populations”. In: *Ann. Math. Statist.* 14.4, pp. 436–440. DOI: 10.1214/aoms/1177731363. URL: https://doi.org/10.1214/aoms/1177731363 (cit. on p. 2).

Du, Ding-Zhu and Frank K Hwang (1999). *Combinatorial Group Testing and Its Applications*. 2nd. WORLD SCIENTIFIC. DOI: 10.1142/4252. URL: https://www.worldscientific.com/doi/abs/10.1142/4252 (cit. on p. 21).

References II

Emmanuel, J C et al. (1988). “Pooling of sera for human immunodeficiency virus (HIV) testing: an economical method for use in developing countries.”. In: *Journal of Clinical Pathology* 41.5, pp. 582–585. ISSN: 0021-9746. DOI: 10.1136/jcp.41.5.582. URL: https://jcp.bmj.com/content/41/5/582 (cit. on p. 3).

Ghosh, Sabyasachi et al. (May 2020). “A Compressed Sensing Approach to Group-testing for COVID-19 Detection”. In: (cit. on p. 20).

Hogan, Catherine A., Malaya K. Sahoo, and Benjamin A. Pinsky (May 2020). “Sample Pooling as a Strategy to Detect Community Transmission of SARS-CoV-2”. In: *JAMA* 323.19, pp. 1967–1969. ISSN: 0098-7484. DOI: 10.1001/jama.2020.5445. URL: https://doi.org/10.1001/jama.2020.5445 (cit. on p. 3).

Mutesa, Leon et al. (2020). “A strategy for finding people infected with SARS-CoV-2: optimizing pooled testing at low prevalence”. In: *medRxiv*. DOI: 10.1101/2020.05.02.20087924. URL: https://www.medrxiv.org/content/early/2020/08/03/2020.05.02.20087924 (cit. on p. 3).

Sobel, Milton and Phyllis A. Groll (1959). “Group Testing To Eliminate Efficiently All Defectives in a Binomial Sample”. In: *Bell System Technical Journal* 38.5, pp. 1179–1252. DOI: 10.1002/j.1538-7305.1959.tb03914.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.1538-7305.1959.tb03914.x (cit. on p. 2).

References III

Sterrett, Andrew (1957). "On the Detection of Defective Members of Large Populations". In: *The Annals of Mathematical Statistics* 28.4, pp. 1033–1036. ISSN: 00034851. URL: http://www.jstor.org/stable/2237067 (cit. on p. 2).

Taylor, Steve M. et al. (2010). "High-Throughput Pooling and Real-Time PCR-Based Strategy for Malaria Detection". In: *Journal of Clinical Microbiology* 48.2, pp. 512–519. ISSN: 0095-1137. DOI: 10.1128/JCM.01800-09. URL: https://jcm.asm.org/content/48/2/512 (cit. on p. 3).

Van, Tam T. et al. (2012). "Pooling Nasopharyngeal/Throat Swab Specimens To Increase Testing Capacity for Influenza Viruses by PCR". In: *Journal of Clinical Microbiology* 50.3, pp. 891–896. ISSN: 0095-1137. DOI: 10.1128/JCM.05631-11. URL: https://jcm.asm.org/content/50/3/891 (cit. on p. 3).

Yelin, Idan et al. (2020). "Evaluation of COVID-19 RT-qPCR test in multi-sample pools". In: *medRxiv*. DOI: 10.1101/2020.03.26.20039438. URL: https://www.medrxiv.org/content/early/2020/03/27/2020.03.26.20039438 (cit. on pp. 3, 7).

Yi, Jirong, Raghu Mudumbai, and Weiyu Xu (Apr. 2020). "Low-Cost and High-Throughput Testing of COVID-19 Viruses and Antibodies via Compressed Sensing: System Concepts and Computational Experiments". In: (cit. on p. 20).