Viral Load Inference in Non-Adaptive Pooled Testing

Mansoor Sheikh[†] **and David Saad**[†]

†Non-linearity and Complexity Research Group, Aston University, Birmingham B4 7ET, United Kingdom

E-mail: d.saad@aston.ac.uk

Abstract. Medical diagnostic testing can be made significantly more efficient using pooled testing protocols. These typically require a sparse infection signal and use either binary or real-valued entries of $\mathcal{O}(1)$. However, existing methods do not allow for inferring viral loads which span many orders of magnitude. We develop a message passing algorithm coupled with a PCR (Polymerase Chain Reaction) specific noise function to allow accurate inference of realistic viral load signals. This work is in the non-adaptive setting and could open the possibility of efficient screening where viral load determination is clinically important.

Keywords: Pooled Testing, Message Passing, Noise models

1. Introduction

Typically the infection status of a patient is determined by carrying out a single diagnostic test, which represents a poor use of resource in the low-prevalence case where most tests return negative. A well-studied improvement is the pooled testing concept [1] which allows for the structured mixing of patient samples into groups or pools. By testing these mixtures (rather than the individual samples), the number of diagnostic tests required to determine each patient's infection status can be dramatically reduced. This can be beneficial where shortages of laboratory diagnostic equipment, raw testing materials and qualified staff occur. The two main approaches to pooled testing are the adaptive and non-adaptive protocols. In the former, tests are run sequentially, with information from the previous testing steps informing the next one [1][2]. However during the early spread of a virulent pathogen, laboratories are typically running at capacity and the logistics of adaptive pooled testing are not always feasible. In this paper, we focus on non-adaptive pooled testing which requires only one testing procedure to infer infection status.

We consider the problem of recovering an unknown N-dimensional vector $\mathbf{x}$ representing the diagnostic status of N patients where component values can be either binary or real. We will introduce the concept using binary values. Combinations of samples are pooled according to the $M \times N$ pooling/measurement matrix, $\boldsymbol{F}$ (see Fig. 1) where each row specifies which patient samples are included in each test. Matrix $\boldsymbol{F}$ can be constructed as either random or structured and is a known quantity in the inference problem. Each row of $\boldsymbol{F}$ can be thought of as probing/examining the unknown signal vector by taking a linear projection via a physical pooling of patient samples. The M results are output as an M-dimensional vector, $\mathbf{y}$ and the

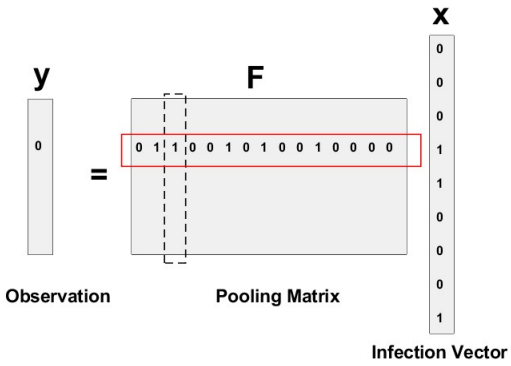

Figure 1. The inverse problem is to infer the $N \times 1$ infection vector, $\mathbf{x}$, given the $M \times 1$ observation vector, $\mathbf{y}$, and the $M \times N$ pooling matrix, $\boldsymbol{F}$.

aim is to solve the inverse problem of inferring $\mathbf{x}$ from known $\boldsymbol{F}$ and $\mathbf{y}$. A smaller measurement rate $\alpha = M/N$ corresponds to a more efficient pooled testing setup. The judicious design of an efficient pooling regime helps minimize α. Typical approaches for real-valued variables $\mathbf{x}$ use pooling matrices with Gaussian random variable entries. Subsequent improvements include matrix designs from error-correcting codes [3] and from physics-inspired methods of crystal nucleation [4].

Individual infection status is determined from the pooled test measurements via a statistical or combinatorial inference procedure with the method used depending on the infection status represention e.g. binary status (typical in group testing) or real-valued viral loads (compressed sensing). These approaches to group testing have been investigated in the fields of computer science [5], statistics [6][7], error-correcting codes [3] and statistical physics [4][8][9][10]. Group testing typically assumes a sparse infection signal i.e. low disease prevalence.

In some applications, the viral load can range from approximately $10^2 - 10^9$ copies per mL e.g. in the case of SARS-CoV-2 [11]. This makes inference problematic, especially for inferring real-valued variables, and, to our knowledge, there is a lack of work dealing with this case. The problem stems from each pooled measurement containing a linear combination of zero, low, medium and high real values. Informally, the lower values will necessarily be "drowned out" by the high values in the pooling process. In this work, we do not assume that the signal entries are $\sim \mathcal{O}(1)$. First we ask the natural question of whether it is clinically relevant to ascertain the viral load of a sample rather than a simple absence or presence of the virus; the answer depends on the application:

- *Viral load is clinically relevant but the efficiencies of pooled testing are not required.* For chronic diseases such as HIV, a lower viral load means higher life expectancy [12]. Hence the success of treatment regimes is measured by viral load. For a previously infected patient, knowing the binary presence or absence does not matter because the virus will be there in some amount. In the case of HIV, the efficiencies of group testing are typically not required (unless it is done as community screening e.g. [13]).
- *Widespread screening required but mostly the presence or absence of the disease is*

relevant: In some cases knowing the viral load is less likely to affect the treatment or isolation regime (such as COVID or Hepatitis C). Here low viral loads mean early infection stages but are less likely to affect the clinical decision. A possible *caveat* to this is evidence that patients with high viral load are more efficient at spreading infection [14] [15]. Identifying and quarantining these individuals will have an outsized effect on viral spread.

- *Both viral load and efficient screening are relevant*: Pooled testing can be used to estimate the amount of contaminants in food. Here it does matter how much salmonella there is in your chicken. There is also a clear commercial benefit for food companies to gain efficiencies from pooled testing.
- *Inferring the stage of outbreak*: The distribution of C_t values in a population is related to the stage of the viral outbreak [16][17]. By inferring C_t values, the method described in the current paper provides a resource-efficient method of estimating the stage of infection and potentially the viral reproduction number, R_0.

A naive approach would be to run a compressed sensing algorithm to recover the viral load signal which has a high dynamic range. Compressed sensing is a theoretically and empirically accurate signal recovery scheme [18][6][10] where the typical error is orders of magnitude smaller than the signal magnitude. However, this results in small signal components becoming indistinguishable from the noise. Synthetic studies conventionally sample the non-zero viral loads from a uniform distribution e.g. $x \sim \mathcal{U}(2^0, 2^{15})$ in [7]. Here approximately 90% of the samples lie in the range $(2^{12}, 2^{15})$, which is above the typical noise level in compressing sensing algorithms, leading to flattering accuracy. The claimed high dynamic range is *probably* not that high. This argument is applicable to a lesser extent in [19] where viral loads are sampled *uniformly* from $[0, 1000]$ but discretization results in 70% of infected samples being in the mid $[300, 700)$ and high $[700, 1000]$ categories. In the current paper, we sample uniformly from a logarithmic range i.e. equal number of samples from each of $2^0, 2^1, 2^2, \ldots$ which is more realistic and relevant. This ensures there is truly a high dynamic signal range but results in a materially harder problem.

2. Model

2.1. Standard combination protocols

There are two typical combination protocols associated with the inverse problem of Fig. 1:

(i) Standard matrix multiplication of $\boldsymbol{F}$ and $\mathbf{x}$ e.g. [4]

$$\tilde{y}_\mu = \sum_{i=1}^{N} F_{\mu i} x_i + \xi_\mu \tag{2.1}$$

where ξ_μ represents Gaussian noise associated with test μ, x_i the individual loads, $F_{\mu i}$ the mixing matrix and $\tilde{y}_\mu$ the noisy test result (a corrupted version of $\mathbf{y}$). This is termed the *linear estimation problem* [20]

(ii) The binary testing regime uses a logical sum e.g.[8]. If one patient is infected, the test output is infected.

$$\tilde{y}_\mu = C\left(\bigvee_{i=1}^{N} F_{\mu i}\ x_i\right) \tag{2.2}$$

where x_i the individual presence/absence of the disease, $\tilde{y}_\mu$ the noisy test result and $C(\ldots)$ is a probabilistic function incorporating measurement noise such as false positive and false negative rates. Since each patient participates in multiple tests (so-called overlapping tests), the accuracy achieved can improve upon the device error settings [8].

2.2. PCR specific notation

In this paper, we focus on amplification methods such those used in the Polymerase Chain Reaction (PCR) device. The initial patient sample is repeatedly heated and cooled to encourage a doubling of the viral RNA (see Fig. 2 for a schematic representation). Hence the initial viral load is amplified. A marker is added which fluoresces when attached to a specific section of the virus. Once sufficient fluorescence is detected with an optical device, the cycling is stopped and the resulting cycle number recorded. If no fluorescence is detected after an upper limit of cycles, typically 40, it is determined that no virus was originally present. The upper cycle threshold (C_t) limit is determined by the *limit of detection* parameter of the PCR device. A C_t value of 20 typically signifies a high viral load. We now define the notation related to the PCR protocol.

- θ = threshold for detection of fluorescence. This is typically measured in number of viral copies per mL of transport media and is assumed constant for a given PCR device.
- a_i = initial viral load (number of viral copies per mL). The higher the initial viral load a_i, the lower the number of amplification cycles needed to detect the virus.
- t_i = number of amplification cycles required to detect fluorescence if sample i is tested individually. Values of t_i are read off as integers rather than the "exact" real valued solution to $\theta = a_i 2^{t_i}$.
- t_μ^y = number of amplification cycles required to detect fluorescence for the pooled test μ.

Given the nature of the PCR doubling cycle, the linear projections described in (2.1) no longer hold. This is because the viral loads are averaged in the pooling process but these viral loads are only observed via the integer cycle number. A new combination protocol is formulated in the next section.

2.3. Simple mixing example

To gain intuition for the idiosyncrasy of PCR mixing, consider two patients i and j where $\theta = a_i 2^{t_i^\theta}$ and $\theta = a_j 2^{t_j^\theta}$ represent each patient being tested individually (note that t^θ are not integers). The aim is to recover t_i and t_j from the possibly noisy measurement vector $\tilde{\mathbf{y}}$. If these samples are combined in an equal ratio in test μ, the resulting viral load, which is never directly observed, is $\frac{1}{2}(a_i + a_j) = \frac{\theta}{2}(2^{-t_i^\theta} + 2^{-t_j^\theta})$. The corresponding measurement cycle number, which we label t_μ^y is given by the

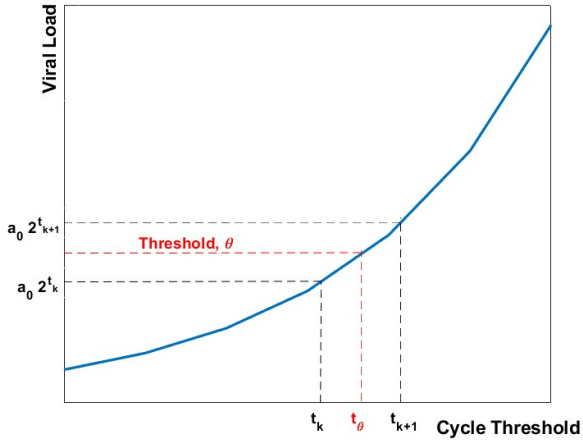

Figure 2. This schematic plot shows the doubling process central to PCR devices. Since detection is at integer values, the observed C_t reading is "rounded-up" from the exact solution to $a_0 2^{t^\theta} = \theta$ where a_0 is a random variable which represents the initial viral load of the sample (notice that t^θ is not an integer). We do not observe a_0 directly but only via an integer valued $\lceil t^\theta \rceil$. Reading off from curve $a_0 2^{t_k} < \theta \leq a_0 2^{t_{k+1}}$. Re-arranging $\theta 2^{-t_{k+1}} \leq a_0 < \theta 2^{-t_k}$.

relationship $2^{-t_\mu^\theta} = \frac{1}{2}(2^{-t_i^\theta} + 2^{-t_j^\theta})$ where the θ parameters cancel. To illustrate numerically, suppose the initial viral loads are $a_i = 2^5 = 32$ and $a_j = 2^6 = 64$ and the detection threshold $\theta = 2^9 = 512$. We find $t_i^\theta = t_i = 4$ and $t_j^\theta = t_j = 3$ noting the inverse relationship between t_i^θ and a_i. For samples pooled in an equal ratio, the viral load per unit volume is $\frac{1}{2}(32+64) = 48$ leading to $t_\mu^\theta \approx 3.42$, a non-integer value between t_i^θ and t_j^θ. The PCR device will return a measurement C_t value of $\lceil t_\mu^\theta \rceil = 4$. This measurement discretization makes the inference problem more difficult and less accurate.

2.4. General mixing example

If equal quantities of K samples are pooled (where $K < N$), the corresponding mixing relationship is

$$2^{-t_\mu^\theta} = \frac{1}{K} \sum_{k=1}^{K} 2^{-t_{\mu k}^\theta} \tag{2.3}$$

where $t_{\mu k}^\theta$ represents the real-valued threshold of element k in the mix and t_μ^θ the real-valued threshold for test μ.

2.5. Signal sparsity assumption

In traditional testing, we perform N diagnostic tests for N patients. Attempts to reduce the number of tests i.e. $M < N$ will lead to an ill-defined system of equations in Fig. 1 with an infinite number of solutions. The constraint required to solve the problem typically relies on a sparsity assumption (in some suitable basis) i.e. a majority of zero entries corresponding to non-infected patients. In the present case, uninfected patients correspond to a cycle number equal to the upper limit. In

typical compressed sensing scenarios, a sparse prior is defined via the signal density ρ using a Bernoulli-Gauss distribution [9] such as $p(\mathbf{t}) = \prod_{i=1}^{N} \left[(1-\rho)\ \delta(t_i) + \rho\ \mathcal{N}(\mu, \sigma^2)\right]$ and i.i.d. signal components. In this paper, the variables used are all integers, corresponding to the PCR cycles where infection can be first detected. They range from the lowest number of cycles L to the highest U. In the absence of additional information we assume the non-zero signal entries take uniformly distributed values in the state space $\mathcal{S} \in \{L, L+1, \ldots, U\}$, according to

$$p(\mathbf{t}) = \prod_{i=1}^{N} \{(1-\rho)\delta_{t_i,U} + \rho\ \Theta[t_i - L]\ \Theta[(U-1) - t_i]\} \tag{2.4}$$

where $\Theta(x)$ is the Heaviside step function and U represents no viral load/not infected.

An alternative approach, used in [5], is to discretize the entire signal range $\mathcal{S} \in \{L, \ldots, U\}$ into low, medium, high and non-infected ranges corresponding to clinically relevant ranges of low, mild and high infection status. This simplification simplifies the inference problem at the expense of accuracy.

3. Message passing

We aim to recover the N-dimensional vector $\mathbf{t}$ where each component $t_i \in \mathcal{S}$. The viral load variables only participate through the mixing relationship described in Sec. 2.4. The inference problem can be represented by a bipartite factor graph (see Fig. 3) and since the graph is sparsely connected, inference can be efficiently achieved using a message passing algorithm, whereby conditional probabilities are iteratively exchanged between factors and variables until they converge to provide pseudo marginal posterior probabilities for the individual variables. Message passing methods are exact on trees but offer approximate solutions on loopy graphs [21]. Their use in the context of group testing has been studied previously but mostly in the case of binary infection status to our knowledge [8, 22]. While preparing the manuscript, we came across the paper [5] which utilizes a message passing algorithm for inference of a real-valued signal. However, there is a distinct difference between our approach and the one of [5], that is based on the iterative removal of values found and uses a very small number of categories, arguably making the problem easier.

Messages exchanged between factors and variables represent a closed set of equations of messages from factors to variables $p(f_\mu|t_i) \equiv m_{\mu\to i}(t_i)$ and from variables to factors $p(t_i|\{\mathbf{f}\}_{\setminus\mu}) \equiv m_{i\to\mu}(t_i)$, detailed below.

3.1. Factor-to-variable messages (PCR noise)

Here we denote the noise model as $\phi(\mathbf{t}, \mathbf{t}^y)$ where $\mathbf{t} \in \{t_1, t_2, \ldots, t_K\}$. The mechanism needed for the discretization inherent in the PCR measurement process is developed in Sec. 4.

$$m_{\mu\to i}^{(\tau+1)}(t_i) \propto \sum_{\mathbf{t}\setminus t_i} \phi(\mathbf{t}, \mathbf{t}^y) \prod_{j\in\partial\mu\setminus i}^{N} m_{j\to\mu}^{(\tau)}(t_j) \tag{3.1}$$

The superscript τ denotes the iteration step and the notation $\partial\mu\setminus i$ refers to all variables connected to factor μ except i. Each message is an $|\mathcal{S}|$-dimensional column

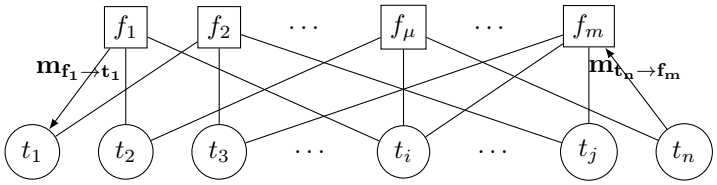

Figure 3. Random variables (circles) represent each of the N patients tested. Each factor node (squares) represents the compatibility of the test/measurement μ with the relevant patient variable values. The degree of the factors nodes is K and of the variable nodes is L

vector with entries being real numbers between 0 and 1 representing the state probability. The summation is over all possible states of the vector $\mathbf{t}$ excluding t_i. This is the main computational bottleneck and does not scale well with $|\mathcal{S}|$ (state space size) or K (factor node degree). This difficulty is reduced in [5] by assuming a state space of none, low, medium and high viral loads. For real-valued signals, the messages can be approximated by their means and variances leading to approximate message passing protocols [10, 18].

3.2. Variable-to-factor messages

$$m_{i\to\mu}^{(\tau+1)}(t_i) \propto \Big\{(1-\rho)\delta_{t_i,U} + \rho\ \Theta[t_i - L]\ \Theta[(U-1) - t_i]\Big\} \prod_{\gamma\in\partial i\backslash\mu} m_{\gamma\to i}^{(\tau)}(t_i) \qquad (3.2)$$

where δ_{ij} is the Kronecker delta function and the notation $\partial i\backslash\mu$ refers to all factors connected to variable i except μ. For both factor-to-variable and variable-to-factor messages, the constants of proportionality can be calculated by normalisation e.g. $\sum_{t_i\in\mathcal{S}} m_{i\to\mu}(t_i) = 1$.

3.3. Marginal Probabilities

Once the messages (3.1)(3.2) have been iterated to convergence, the marginal probabilities for each variable/patient can be calculated:

$$p(t_i) \propto \Big\{(1-\rho)\delta_{t_i,U} + \rho\ \Theta[t_i - L]\ \Theta[(U-1) - t_i]\Big\} \prod_{\mu\in\partial i} m_{\mu\to i}(t_i) \qquad (3.3)$$

The inferred C_t value corresponds to the message component with highest probability. It is an approximation in our case since the bipartite graph contains loops.

4. Developing the noise model

The two types of noise present in the PCR mechanism are discretization noise and physical measurement noise. We will address the former in this section and the latter in Sec. 5.6. Discretization noise is present since C_t values are recorded as integers rather than decimals and requires careful treatment in the mathematical modelling setup.

4.1. No measurement noise

In the case where the real cycle values are applied (denoted by the θ superscript) Eq. (2.3) holds, and consequently

$$\phi(\mathbf{t}, \mathbf{t}^y) = \prod_{\mu=1}^{M} \delta\left(\frac{1}{K}\sum_{k=1}^{K} 2^{t_\mu^\theta - t_{\mu k}^\theta} - 1\right) \tag{4.1}$$

However, discretization to the next higher integer cycle value creates a difference between the test cycle integer value and those of the individual samples. The modelling of this difference is a challenge and an appropriate noise model has to be used.

4.2. Gaussian measurement noise

One of the simplest models to accommodate these differences is to assume measurement errors follow a Gaussian distribution incorporating (2.3) and the discretization (see [7]). We will replace the "true" but unknown real cycle values t_μ^θ and $t_{\mu k}^\theta$ by the integer variables t_μ^y and $t_{\mu k}$, respectively. We expect that $t_\mu^y = \lceil t_\mu^\theta \rceil$ and $t_{\mu k} = \lceil t_{\mu k}^\theta \rceil$.

$$\phi(\mathbf{t}, \mathbf{t}^y) = \prod_{\mu=1}^{M} \frac{1}{\sqrt{2\pi\Delta_\mu}} \exp\left[-\frac{1}{2\Delta_\mu}\left(\frac{1}{K}\sum_{k=1}^{K} 2^{t_\mu^y - t_{\mu k}} - 1\right)^2\right] \tag{4.2}$$

where Δ_μ is the variance in the noise measurement.

We will consider two different noise distributions $\phi(\mathbf{t}, \mathbf{t}^y)$ to "filter out" unreasonable combinations of $\mathbf{t}$ and $\mathbf{t}^y$ from the factor-to-variable messages in (3.1). The first is a simple binary function in Sec. 4.3 and the second uses the overlap between distributions to weight message products in Sec. 4.4.

4.3. Step-function distribution

First consider the measure d inspired by (2.3) but using integer rather than real values.

$$d = \left(\frac{1}{K}\sum_{k=1}^{K} 2^{-t_k}\right) - 2^{-t^y} = X - Y \tag{4.3}$$

where we will use short-hands $X \equiv \frac{1}{K}\sum_{k=1}^{K} 2^{-t_k}$ and $Y \equiv 2^{-t^y}$ which are fixed for a given combination. The uncertainty in both random variables X and Y can be represented by uniform distributions $\mathcal{U}(X, 2X)$ and $\mathcal{U}(Y, 2Y)$ leading to an inequality:

$$\frac{1}{2}2^{-t^y} \leq \frac{1}{K}\sum_{k=1}^{K} 2^{-t_k} \leq 2 \cdot 2^{-t^y} \tag{4.4}$$

Hence our first approximation for $\phi(\mathbf{t}, \mathbf{t}^y)$ can be written:

$$\phi(\mathbf{t}, \mathbf{t}^y) = \begin{cases} 1 & \text{if } -\frac{1}{2}2^{-t^y} \leq d \leq 2^{-t^y} \\ 0 & \text{otherwise.} \end{cases} \tag{4.5}$$

Notice this function does not need to be normalized since the messages are normalized in the message passing iterations. Clearly $d = 0$ is in the interval as expected. The message passing summation (3.1) includes all combinations of $\{t_k\}$ for a given t_y and the role of $\phi(\mathbf{t}, \mathbf{t}^y)$ is to exclude implausible combinations. This simple binary approach has the benefit of materially reducing the number of summands in each message passing iteration speeding up implementation.

4.4. Distribution Overlap

Rather than a binary function, consider a function $\phi(\mathbf{t}, \mathbf{t}^y)$ which weights different combinations of $\{t_k\}$ and t_y depending on the overlap of the distributions $\mathcal{U}(X, 2X)$ and $\mathcal{U}(Y, 2Y)$. For the case $X < Y$:

$$\phi(\mathbf{t}, \mathbf{t}^y) = \begin{cases} 0 & \text{if } X \leq \frac{1}{2}Y \\ \frac{2X-Y}{Y} & \text{if } \frac{1}{2}Y \leq X \leq Y \\ \frac{2Y-X}{Y} & \text{if } Y \leq X \leq 2Y \\ 0 & \text{if } X > 2Y \end{cases} \tag{4.6}$$

This noise distribution, which has the same support as (4.4), is used in the numerical experiments of Sec. 5.

4.5. Further refinement

Using a uniform distribution for 2^{-t^y} is plausible but the distribution for $X = \frac{1}{K}\sum_k 2^{-t_k}$ should properly account for each t_k chosen at random from $\{L, \ldots, U\}$. The probability density function has been derived in [23][24]. The cumulative distribution function for X, derived in [25], can be used to calculate the overlap with $\mathcal{U}(2^{-t^y}, 2 \cdot 2^{-t^y})$. This calculation is not implemented due to the increased computational cost required due to the nested summations.

4.6. Alternative treatments

Previous research has considered the PCR noise function but primarily in the case of binary variables (infected/non-infected). The traditional noise measures of false positive and false negative rates are comprehensively treated in [8] where it is shown that pooled testing can overcome technical device limitations since each patient participates in multiple tests. Noise in C_t values and fluoresence thresholds are dealt with in [7] but with a different signal reconstruction algorithm and the discretization effect is modelled as Gaussian rather than uniform. We have avoided the need for calibration of the threshold θ and mapping between amplified viral load and fluoresence in [7] with our functional form of $\phi(\mathbf{t}, \mathbf{t}^y)$. An alternative approach is to learn the translation function between viral loads and discrete measurement classes using supervised learning methods such as a neural network [5]. While our approach splits this up into mixing (2.3) and noise probability functions, [5] may be able to learn from previous PCR results *if the ground truth properties are known.*

5. Numerical Experiments

5.1. Implementation

The problem is defined by initializing parameters ρ, M, N, K and state space $\mathcal{S}$. Of course, the signal sparsity ρ is not *a priori* known for a given batch of samples but could be estimated by either the Expectation-Maximization [8][9] or the Expectation Propagation [26] methods. An approximate but practical approach is to take a recent average disease prevalence value as an estimate of ρ. This parameter is used to set the number of observations M in the pooling matrix.

We now carefully describe the protocol for generating the ground truth signal $\mathbf{t}$ in our numerical experiments. This is important as it influences our subsequent choice of noise function boundary. *Decimal* values are sampled for infected and non-infected patients from uniform distributions $\mathcal{U}(L-1, U-1)$ and $\mathcal{U}(U-1, U)$ respectively and combined using (2.4). The *decimal* measurement vector t_y^θ is calculated *in silico* using the PCR mixing protocol (2.3). Finally an *integer* measurement value is obtained by rounding up t_y^θ to the nearest integer. Additional noise is added (see Sec. 5.6) in some scenarios as described below. The discretized version of $\mathbf{t}$ is stored for calibration purposes once the inference is complete. A different (unphysical) protocol could be considered which starts from *integer* values of infected patients. This would reduce uncertainty in the noise function thus providing more flattering accuracy statistics but is not used in our experiments since it unrealistic.

The random binary measurement matrix $\boldsymbol{F}$ defines which samples are included in each test and is typically defined by the degree distribution of the variable and factor nodes of the bipartite graph (Fig. 3). Choices include random-random or random-Poisson [22]. In this paper, we follow [8] in using random-random matrices such that each row has K non-zero entries and each column has $L = KN/M$ non-zero entries. In other words, the factor nodes have a constant degree K and the variable nodes a constant degree L. The values used in our setup result in a sparse $\boldsymbol{F}$ matrix. Our message passing algorithm could also be combined with the structured matrix approach [4, 9, 27].

Given knowledge of $\boldsymbol{F}$ and $\mathbf{y}$, the message passing equations (3.1), (3.2) are iterated until convergence. The message passing equations search for a fixed point to the dynamical system of messages, from which posterior variable values can be inferred using (3.3). Typical of iterative problems, our implementation utilises a damping factor (in our case, set to 0.01) to help convergence. In addition, the sum of squared differences between all messages is chosen as a metric to determine convergence. The threshold was chosen empirically as 10^{-5} for the noiseless cases. This synthetic setup can therefore be used to test the accuracy of signal recovery.

5.2. Easy Algorithm

In problems where only binary infection status is considered, *sure variables* of certain states can be determined using simple logical arguments without resorting to more complex inference methods. These are termed *Combinatorial Basis Pursuit (CBP)* or *Combinatorial Orthogonal Matching Pursuit (COMP)* in computer science implementations of group testing [28] and are described in an *Easy Algorithm* [22]. Briefly, they identify certain (sure) negative patients since they participate in a negative test result. Further, sure positive samples can be found as those who participate in positive tests alongside sure negative samples. These variables are then

removed reducing the problem size. One drawback of this approach is the inability to infer *undetermined variables* [22], the inability to use prior knowledge and to exploit the power of probabilistic inference in the inevitable presence of measurement noise.

In the present case, where the state space $\mathcal{S}$ represents C_t values, a non-infected test measurement result, $t_\mu^y = U$, does not guarantee all samples are non-infected (even in the noiseless case) e.g. true C_t values of $\{U-1, U, U, U, U\}$ will result in $t_\mu^y = U$. This is due to the peculiarity of the PCR mixing protocol (2.3) and the discretization of C_t values. The issue can be seen clearly for the low viral load values (equivalently high C_t values) in Fig. 4. Hence the *Easy Algorithm* cannot be used in our work, either as a pre-processing step or as a comparison baseline for our message passing results. In fact it is arguably impractical to any scenario that includes a measurement noise.

5.3. Uncertain Inference

Recall that each message consists of $|\mathcal{S}|$ components representing the probability of each state from L to U. To be clear, each state of $\mathcal{S}$ represents a doubling of the viral load from the preceeding state on a logarithmic scale (base 2). Our Maximum A Posteriori *(MAP)* inference protocol uses the highest probability state as our diagnosis. However, for some samples, two states could have similar high probabilities implying uncertain inference. A *computational* improvement could be envisaged, using decimation, where, upon convergence, the less ambiguous results are fixed to a standard basis vector e.g. $(0, 1, 0, \ldots, 0)^T$ and the message passing algorithm continues with a smaller convergence threshold. Note, the "certain" samples are simply the complement of the uncertain samples. This intervention in the algorithm dynamics did not meaningfully change the results in our test implementations. A possible *clinical* protocol would be to physically re-test this small number of individuals with ambiguous results, but this has its own operational and medical implications.

5.4. Motivating Example

To illustrate the discretization problem central to PCR testing, Fig. 4 shows the confusion matrix for the noiseless inference problem of $N = 2400$ individual samples, $M = 800$ observations, $K = 6$ patients in each group, $\rho = 0.01$ prevalence and $\mathcal{S} \in \{20, \ldots, 30\}$. As a guideline, the expected number of tests required in adaptive *binary (yes/no only)* Dorfmann tests is approximately $2\sqrt{\rho}N = 480$ (see [22]). Predicting a C_t value one away from its true value is clinically acceptable leading to the formulation of two accuracy measures C_0 and $C_{\pm 1}$ corresponding to a tolerance of zero and ± 1 C_t value, respectively. However, given the imbalanced classes in the signal, this metric is found not to be suitable to our task since simply predicting no infection results in high accuracy.

For the remainder of this paper, we will focus on the clinically relevant metric of sensitivity (the ratio of true positives to positive samples) and use a state space comprising $\mathcal{S} \in \{20, \ldots, 30\}$. The relationship between sensitivity and $|\mathcal{S}|$ is investigated in Fig. 5.

5.5. Accuracy of inference procedure

We carry out numerical experiments with a fixed measurement ratio $\alpha = M/N = 800/2400 = 0.33$ and vary the signal sparsity in the range $\rho = [0.01, 0.05]$ to represent

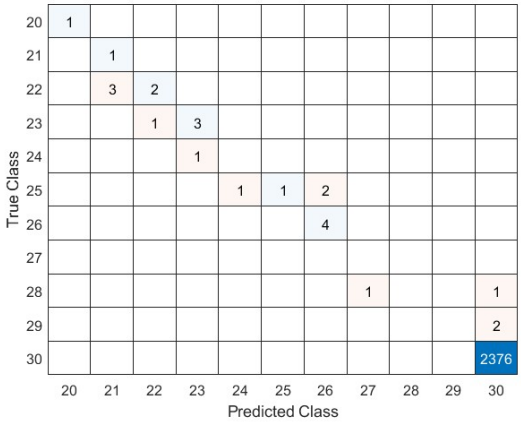

Figure 4. Confusion matrix representing a single problem instance, described in Sec. 5.4, using synthetic data: $M = 800$, $N = 2400$, $K = 6$, $\rho = 0.01$, $\eta = 0$ and $\mathcal{S} \in \{20, \ldots, 30\}$. The noise function $\phi(\mathbf{t}, t_y)$ is taken as described in Sec. 4.4, producing an accuracy of 2388/2400 or 99.5% (compared to 99.0% for a always-negative classifier). The false negative rate of $3/24 = 12.5\%$ reduces as the state space $\mathcal{S}$ increases (see Fig. 5).

typical infectious disease values. 30 simulations are run for each ρ value with C_0 and $C_{\pm 1}$ accuracy, true/false positive/negative measures recorded. The high accuracy values shown in Fig. 6 mask an underlying issue with the false negative rates hence our focus on sensitivity.

5.6. Accuracy versus measurement noise

In previous sections, we have accounted for the discretization noise arising from the PCR doubling mechanism. Physical noise sources, also inherent in recording t^y values, include fluorescence not being distinguished from background levels e.g. light leaks into the sample well and imperfect doubling during the amplification phase e.g. $(1+q)^t$ where $q \in (0, 1)$ rather than 2^t [7]. These tend to lower and raise the measured C_t value from its true reading respectively. In the absence of accurate statistics, we naively assume under- and over- C_t estimation is equiprobable leading to a noisy measurement $\tilde{t}^y$:

$$p(\tilde{t}^y) = (1-\eta)\ \delta(\tilde{t}^y - t^y) + \frac{1}{2}\eta\ \delta\left[(\tilde{t}^y - t^y) - 1\right] + \frac{1}{2}\eta\ \delta\left[(\tilde{t}^y - t^y) + 1\right] \tag{5.1}$$

where $\eta \ll 1$. Setting $\eta = 0$ represents no physical measurement noise (which was the case for our previous experiments). Equation (5.1) is used to add noise to our synthetically generated signals. After convergence of our message passing equations, the inferred C_t values are converted into binary variables where negative corresponds to $C_t = U$ and positive otherwise. The resulting sensitivity values, plotted in Fig. 7, assess how robust the inference is to increasing η. The presence of overlapping tests i.e. each patient participating in multiple tests ($L = KN/M > 1$) was considered in [8] and was found to overcome typical noise levels.

In addition to device-specific measurement noise, modelled by (5.1), laboratory sample handling errors can occur [29]. These include cross-contamination of samples

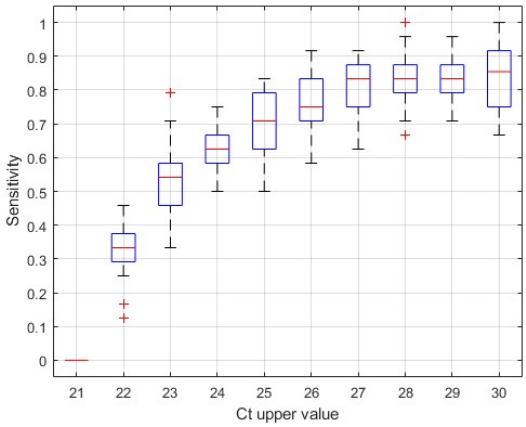

Figure 5. Plot of Sensitivity (True Positive Rate = TP/P) against the largest C_t value. The lower C_t value is fixed at 20. The intuition from our motivating example in Fig. 4 is false negatives occur for C_t values close to the maximum. This plot confirms false negatives decrease as the range size become larger the number of values per variable is higher. The zero sensitivity value at $C_t = 21$ corresponds to all 24 positive samples being falsely classified as negative. Parameters $\rho = 0.01$, $\alpha = M/N = 800/2400$ are fixed. Changing a true to a false positive results in a sensitivity change of $1/24 \approx 4\%$. Simulations are run 30 times. Red lines display the median value, the boxes cover the interquartile range and the whiskers show the extreme values.

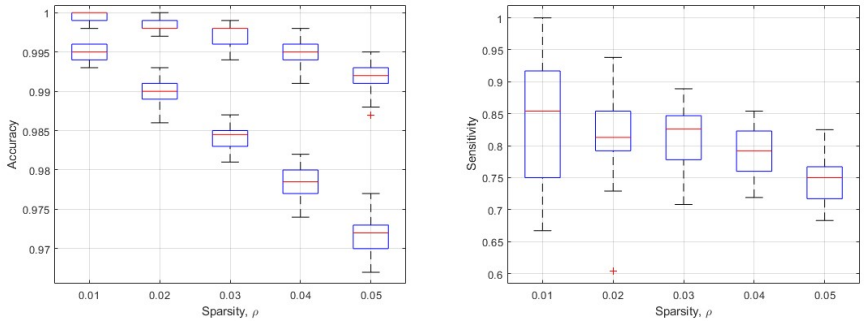

Figure 6. Left: Plot of C_0 (lower) and $C_{\pm 1}$ (upper) accuracy measures against signal sparsity ρ. Right: Plot of Sensitivity against ρ. Note that specificity (True Negative Rate = TN/N) is ≥ 0.999 for $\rho \in [0.01, 0.05]$ so is not plotted. The state space is $\mathcal{S} \in \{20, \ldots, 30\}$. Simulations are run 30 times and the red lines display the median value, the boxes cover the interquartile range, the whiskers show the extreme values and outliers are shown with red '+' symbols. The signal sparsity/prevalence range (x-axis) is chosen to represent typical infectious diseases values. The measurement ratio α is constant at $M/N = 800/2400$ and no measurement noise is added ($\eta = 0$, see discussion in Sec. 5.6). These plots correspond to the experiment described in Sec. 5.5.

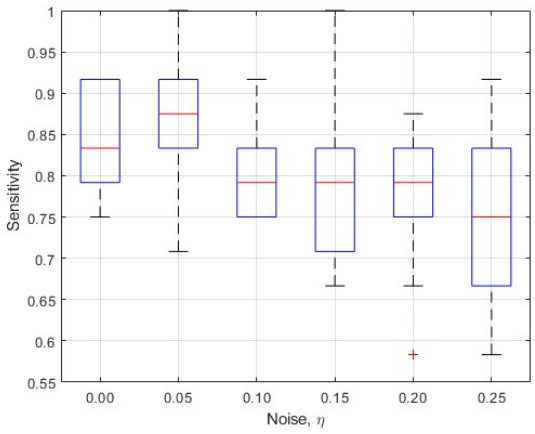

Figure 7. Sensitivity against noise level η defined in (5.1). This corresponds to experiment described in Sec. 5.6. Parameters $\rho = 0.01, \alpha = M/N = 800/2400$ are fixed. Simulations are run 30 times. The necessity of more iterations for the noisy signals is well known [22] so it was necessary to relax the convergence threshold to 10^{-3}.

and errors pipetting samples into physical groups/pools. This source of error is not accounted for in our model.

6. Discussion

Pooled testing was originally designed for the efficient diagnosis of infectious diseases [1]. There are numerous explanations for the lack of widespread adoption in public health including perceived complexity, requirement for sophisticated laboratory management or conflicting economic incentives. By combining pooled testing efficiencies with accurate estimation of a realistic range of viral loads, the methods presented here may lead to wider adoption into applications such as food contaminant testing and community screening for infectious diseases.

It is known that the viral load present in infectious disease testing samples can vary by many orders of magnitude [11]. This concept has typically not been taken into account in theoretical investigations of pooled testing. We estimate the viral load by mapping the inference task to a message passing problem where the factor nodes incorporate the specific PCR mixing protocol. Numerical experiments explicitly highlight the source of the error originating from mixing patient samples and we have dealt with this through a modified noise function $\phi(\mathbf{t}, \mathbf{t}^y)$. This filter, used in Sec. 4.3, approximates the distributions of random variables $X \equiv \frac{1}{K} \sum_{k=1}^{K} 2^{-t_k}$ and $Y \equiv 2^{-t^y}$ as step functions with the overlap providing weights in the message passing equations. For completeness, we note that for applications where the viral load actually has a narrow range of values, the well-developed theory of compressed sensing [9][10] can be applied to infer signal values, accurately and efficiently, using either random or structured pooling matrices.

The main focus of future work is to improve the scalability of algorithm. While real-valued signals can be approximated via the first two moments of the message

distributions, the approximation for our integer-valued state space problem is not clear. The resulting summation in the factor-to-variable messages (3.1) does not scale well with $|\mathcal{S}|$ (increasing C_t range) or K (number of patients in each pool). Our choice of noise function mitigates this effect by materially reducing the number of summands in the factor-to-variable messages. An alternative approach is to translate the integer C_t values to broad classes of infection status e.g. [5]. Further code efficiencies or parallelization may also help.

Other avenues of future work include a wet laboratory experiment to test our modelling assumptions. The methods described in this paper can be applied to other amplification protocols such as Loop-mediated isothermal amplification (LAMP).

7. Acknowledgments

This work was funded by the Engineering and Physical Sciences Research Council through grant EP/W015412/1. The authors would like Professor Andrew Beggs for an earlier collaboration on group testing for the SARS-CoV-2 diagnosis.

References

[1] R Dorfman. The detection of defective members of large populations. *Ann. Math. Stat.*, 14(4):436–440, 1943.
[2] M Mézard and C Toninelli. Group testing with random pools: Optimal two-stage algorithms. *IEEE T Inform Theory*, 57(3):1736–1745, 2011.
[3] N Shental et al. Efficient high-throughput sars-cov-2 testing to detect asymptomatic carriers. *Science advances*, 6(37):eabc5961, 2020.
[4] P Zhang et al. Non-adaptive pooling strategies for detection of rare faulty items. In *2013 IEEE Int Conf Comm*, pages 1409–1414. IEEE, 2013.
[5] Ben-Knaan et al. Recovery of noisy pooled tests via learned factor graphs with application to covid-19 testing. In *ICASSP 2022-2022 Int Conf Acoust Spee*, pages 4518–4522. IEEE, 2022.
[6] D L Donoho et al. Message-passing algorithms for compressed sensing. *P Natl Acad Sci*, 106(45):18914–18919, 2009.
[7] S Ghosh et al. A compressed sensing approach to pooled rt-pcr testing for covid-19 detection. *IEEE Open Journal of Signal Processing*, 2:248–264, 2021.
[8] A Sakata. Bayesian inference of infected patients in group testing with prevalence estimation. *J. Phys. Soc. Jpn.*, 89(8):084001, 2020.
[9] F Krzakala et al. Probabilistic reconstruction in compressed sensing: algorithms, phase diagrams, and threshold achieving matrices. *J. Stat. Mech. - Theory E.*, 2012(08):P08009, 2012.
[10] F Krzakala et al. Statistical-physics-based reconstruction in compressed sensing. *Phys. Rev. X*, 2(2):021005, 2012.
[11] Y Pan et al. Viral load of sars-cov-2 in clinical samples. *Lancet Infect Dis*, 20(4):411–412, 2020.
[12] J Kiplagat and A Justice. Hiv viral suppression is key to healthy longevity. *Lancet HIV*, 9(10):e672–e673, 2022.
[13] M Krajden et al. Pooled nucleic acid testing increases the diagnostic yield of acute hiv infections in a high-risk population compared to 3rd and 4th generation hiv enzyme immunoassays. *J. Clin. Virol.*, 61(1):132–137, 2014.
[14] P M Beldomenico. Do superspreaders generate new superspreaders? a hypothesis to explain the propagation pattern of covid-19. *Int. J. Infect. Dis.*, 96:461–463, 2020.
[15] Y Liu et al. Viral dynamics in mild and severe cases of covid-19. *Lancet Infect Dis*, 20(6):656–657, 2020.
[16] J A Hay et al. Estimating epidemiologic dynamics from cross-sectional viral load distributions. *Science*, 373(6552):eabh0635, 2021.
[17] A Singanayagam et al. Community transmission and viral load kinetics of the sars-cov-2 delta (b. 1.617. 2) variant in vaccinated and unvaccinated individuals in the uk: a prospective, longitudinal, cohort study. *Lancet Infect Dis*, 22(2):183–195, 2022.

[18] Y. Kabashima. A statistical mechanical approach to cdma multiuser detection: propagating beliefs in a dense graph. In *IEEE International Symposium on Information Theory, 2003. Proceedings.*, pages 329–329, 2003.

[19] A Cohen et al. Multi-level group testing with application to one-shot pooled covid-19 tests. In *ICASSP 2021-2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1030–1034. IEEE, 2021.

[20] S Rangan. Generalized approximate message passing for estimation with random linear mixing. In *2011 IEEE International Symposium on Information Theory Proceedings*, pages 2168–2172. IEEE, 2011.

[21] M Opper and D Saad, editors. *Advanced mean field methods: theory and practice*. Neural Information Processing. MIT, February 2001.

[22] M Mézard et al. Group testing with random pools: Phase transitions and optimal strategy. *J Stat Phys*, 131(5):783–801, 2008.

[23] D M Bradley and R C Gupta. On the distribution of the sum of n non-identically distributed uniform random variables. *Ann I Stat Math*, 54:689–700, 2002.

[24] S M Sadooghi-Alvandi et al. On the distribution of the sum of independent uniform random variables. *Stat Pap*, 50(1):171–175, 2009.

[25] A Buonocore et al. A note on the sum of uniform random variables. *Stat Probabil Lett*, 79(19):2092–2097, 2009.

[26] A Braunstein et al. Compressed sensing reconstruction using expectation propagation. *J. Phys. A: Math. Theor.*, 53(18):184001, 2020.

[27] M C Angelini et al. Compressed sensing with sparse, structured matrices. In *2012 50th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 808–814. IEEE, 2012.

[28] C L Chan et al. Non-adaptive probabilistic group testing with noisy measurements: Near-optimal bounds with efficient algorithms. In *2011 49th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pages 1832–1839. IEEE, 2011.

[29] P Courtney and P G Royall. Using robotics in laboratories during the covid-19 outbreak: A review. *IEEE Robot Autom Mag*, 28(1):28–39, 2021.