3-14-2014

Efficiently Identifying Failures using Quantitative Tests, Matrix-Pooling and the EM-Algorithm

Brett Hanscom
University of Washington - Seattle Campus, bhanscom@uw.edu

Susanne May
University of Washington - Seattle Campus, sjmay@uw.edu

Jim Hughes
University of Washington - Seattle Campus, jphughes@u.washington.edu

Efficiently Identifying Failures using Quantitative Tests, Matrix-Pooling and the EM-Algorithm

Brett S. Hanscom, Susanne May, James P. Hughes

3/13/2014

Abstract

Pooled-testing methods can greatly reduce the number of tests needed to identify failures in a collection of samples. Existing methodology has focused primarily on binary tests, but there is a clear need for improved efficiency when using expensive quantitative tests, such as tests for HIV viral load in resource-limited settings. We propose a matrix-pooling method which, based on pooled-test results, uses the EM algorithm to identify individual samples most likely to be failures. Two hundred datasets for each of a wide range of failure prevalence were simulated to test the method. When the measurement of interest was normally distributed, at a failure prevalence level of 15.6% the EM method yielded a 47.3% reduction in the number of tests needed to identify failures (as compared to testing each specimen individually). These results are somewhat better than the reduction gained by using the Simple Search method (44.9%) previously published by May et al. (2010). However, the EM procedure was able to identify failures in just 2.6 testing rounds, on average, as compared to an average of 19.2 testing rounds required by Simple Search. In settings where the turn-around time for testing services is significant, the reduction in testing rounds provided by the EM method is substantial. Unfortunately the EM method does not perform as well when the measurements of interest are highly skewed, as is often the case with viral load concentrations.

Key words

Pooled testing, matrix pooling, quantitative tests, HIV viral load, EM Algorithm

1 Introduction

Pooled-testing methods can greatly reduce the number of tests needed to identify cases of disease in biological samples, particularly when disease prevalence is low. One drawback associated with these methods is that they typically involve lengthy, iterative procedures requiring long turn-around times. Another limitation is that the majority of existing pooling methods only apply to binary tests, i.e. tests that indicate only the presence or absence of a biological agent. In some instances the question of interest is not whether a substance is present, but whether the amount of that substance is higher that a certain threshold. When monitoring HIV-1 viral load in patients treated with ART, for example, the presence of HIV-1 virus is already known, and the questions is whether the concentration of HIV virus has surpassed a critical threshold. In resource limited settings the cost associated with viral-load tests is very high. If pooled testing methods can sufficiently reduce the costs associated with viral-load monitoring, it may become possible to introduce viral-load monitoring in resource limited settings. Ideally we would like to identify "treatment-failure" cases quickly, without the need to carry out a lengthy pooled-testing algorithm.

Here we develop a pooled-testing method which accounts for the amount of a disease agent present in biological samples, and rapidly identifies specific samples that have surpassed a critical threshold. Our method combines 2-dimensional matrix pooling with the EM algorithm, and iteratively tests individual samples. Simulation studies show that this approach can reduce the number of tests needed to identify failures, and dramatically reduces turn-around time. The method is general and can be applied in any setting where the test of interest yields a continuous measure of concentration.

2 Background

Pooled testing, also known as 'group testing', is a method that has been successfully used to reduce the cost of identifying disease cases (or failures) in a set of individuals (or items). Pooled testing is an intuitive method for saving time and money, and has broad application. The basic idea involves taking small portions of each specimen, mixing them together into one or more pools, and then testing the mixed

pools. Assuming the test is sensitive, if a pool tests negative, all individual specimens in that pool must be negative, and the cost of multiple individual tests has been saved. If a pool test is positive, further testing must be performed to identify positive cases in that pool. Provided that a majority of pools yield negative tests, which will generally be true if prevalence is low, substantial cost savings can be achieved.

Dorfman (1943) first quantified the conditions under which pooled testing is useful in the context of binary (positive/negative) tests. By formulating a simple probability model, Dorfman was able to quantify the expected benefit of pooled testing, as well as to compute optimal pool configurations. Testing each of N individual specimens separately can be thought of as the baseline (expensive) approach, and pooled-testing methods are evaluated by the expected percent reduction in tests T required to identify all cases, i.e. $1 - E[T/N]$. Savings can be substantial, for example Dorfman found that under a prevalence of 1% and (optimal) pool size of 11, we can expect an 80% reduction in the number of tests performed. In general, the potential efficiency gain is larger in populations with lower disease (or failure) prevalence.

Since Dorfman's initial paper, numerous authors have proposed a wide variety of improvements and extensions to the basic group-testing idea. For example Phadarfod and Sudbury (1994) proposed using a matrix-pooling approach whereby specimens are arranged into a two-dimensional matrix, and the groups formed by combining samples from each row and each column are tested. Each specimen residing at the intersection of a positive row and positive column is then tested individually. An important advantage to matrix-pooling designs is that by testing both row and column pools each specimen is effectively tested twice, and thus the probability of false-negative samples can be reduced. Phadarfod and Sudbury (1994) showed that by implementing a simple square-array testing scheme the probability of a false-positive sample can be reduced by more than ten-fold in many practical scenarios.

Nearly all published results regarding pooled testing are based on binary testing. Individual samples either contain or do not contain a certain substance, and likewise a pool of individual samples either contains or does not contain that substance. Binary testing is common in biomedical settings, and the conceptual simplicity of binary tests lends itself well to pooled testing methods. There are instances, however, when we are interested in how much of a compound is present. When testing for lead in lake water, for example, one may expect to find a small amount of lead in any given water sample, but would only be concerned if the amount in any individual sample exceeded a certain threshold. Similarly, in the context of viral-load monitoring, all individuals are expected to have low levels of viral RNA in their

blood, but we are only concerned when viral load becomes too high.

Quantitative tests produce more detailed results than binary tests, and as a result pooled quantitative tests can provide more information than binary tests. Exploiting this idea, May et al. (2010) developed a sequential testing algorithm based on matrix pooled samples for identifying ART failure among HIV patients. The algorithm takes its strength from the fact that if a pool of specimens test positive on a binary test, there is no way to tell how many members of the pool are positive. If an individual from that pool is tested and turns out positive, it is still necessary to test the remaining members of the pool (perhaps by re-pooling them) to determine whether other members are positive as well. On the other hand, if a quantitative test is performed on one member of a positive pool, and the amount of test material found in that sample is enough to explain the amount of material observed in the pool, then no further testing of individuals in that pool is necessary. The search algorithm developed by May et al. (2010) is more efficient than other pooling methods at prevalence levels between about 4% and 25%.

We now propose a statistical approach that uses a matrix-pooling strategy to predict which individual samples are most likely to have surpassed the failure threshold. By making assumptions about the distribution of the target substance, and further using the EM algorithm to estimate key parameters in this distribution, we extend the method reported by May et al. (2010). The objective is to improve efficiency, defined as the percent reduction in the number of tests needed to identify all cases, and also reduce the turn-around time, defined as the number of sequential testing iterations required to identify all cases. This method can be used for any application where quantitative testing is performed. Although we use the term "failure" to indicate a concentration value above a critical threshold, in other settings a high concentration may indicate "success". The methodology is the same regardless of how a threshold breach is labelled.

3 Methods

3.1 Overview

Our approach to testing matrix-pooled specimens is an iterative algorithm that alternates between (1) estimating failure prevalence using Expectation-Maximization (EM), and (2) testing the individual specimens that are most likely to contain failure-level concentrations. Test results from individual (non-pooled)

specimens are then fed back into the EM procedure, prevalence is re-estimated, and further individual specimens are identified for testing. Once no further test candidates are found, the procedure is complete.

3.2 The Model

Starting with n^2 specimens arranged into an $n \times n$ array, we form $2n$ row and column pools, and test each pool for the target substance. For i=1...n (rows) and j=1...n (columns) let $\boldsymbol{Y} = \{y_{ij}\}$ be the unobserved concentration values for each specimen. Let $\boldsymbol{Z} = \{z_{ij}\}$ represent the failure status of each specimen, where $z_{ij} = 1$ if the ijth specimen is a failure and $z_{ij} = 0$ otherwise. Assume the z_{ij}s are iid $Bernoulli(p)$ where p is the (unknown) failure prevalence in the population. Assume that the target quantities are normally distributed conditional on z_{ij},

$$y_{ij}|z_{ij} \sim N(\lambda z_{ij} + \theta(1 - z_{ij}), \sigma_f^2 z_{ij} + \sigma_n^2(1 - z_{ij})) \tag{1}$$

where λ is the mean concentration among failures, θ is the mean concentration among non-failures, σ_f is the standard deviation among failures and σ_n is the standard deviation among normals in the population. Since the z_{ij}s are iid, then the y_{ij}s are also independent. Thus if we let $\boldsymbol{y}$ represent the values y_{ij} of the matrix $\boldsymbol{Y}$ arranged into a single column vector $(y_{11}, y_{12}, ..., y_{nn})^T$, then

$$\boldsymbol{y}|\boldsymbol{Z} \sim MVN_{n^2}(\nu(\boldsymbol{Z}), \boldsymbol{\Gamma}(\boldsymbol{Z})) \tag{2}$$

where

$$\nu(\boldsymbol{Z}) = \begin{pmatrix} \lambda z_{11} + \theta(1 - z_{11}) \\ \lambda z_{12} + \theta(1 - z_{12}) \\ \vdots \\ \lambda z_{nn} + \theta(1 - z_{nn}) \end{pmatrix} \tag{3}$$

and

$$\boldsymbol{\Gamma}(\boldsymbol{Z}) = \begin{pmatrix} \sigma_f^2 z_{11} + \sigma_n^2(1 - z_{11}) & 0 & \dots & 0 \\ 0 & \sigma_f^2 z_{12} + \sigma_n^2(1 - z_{12}) & \dots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \dots & \sigma_f^2 z_{nn} + \sigma_n^2(1 - z_{nn}) \end{pmatrix} \tag{4}$$

We can now represent the row and column pool concentrations as a function of the individual specimen concentrations as follows:

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} = \boldsymbol{A}\boldsymbol{y} + \epsilon \tag{5}$$

where

$$\boldsymbol{A} = \frac{1}{n} \begin{pmatrix} 1 & 1 & \dots & 1 & 0 & 0 & \dots & 0 & \dots & 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & 0 & 1 & 1 & \dots & 1 & \dots & 0 & 0 & \dots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & 0 & \dots & 0 & 0 & 0 & \dots & 0 & \dots & 1 & 1 & \dots & 1 \\ 1 & 0 & \dots & 0 & 1 & 0 & \dots & 0 & \dots & 1 & 0 & \dots & 0 \\ 0 & 1 & \dots & 0 & 0 & 1 & \dots & 0 & \dots & 0 & 1 & \dots & 0 \\ \vdots & & & & & & & & & & & & \vdots \\ 0 & 0 & \dots & 1 & 0 & 0 & \dots & 1 & \dots & 0 & 0 & \dots & 1 \end{pmatrix} \tag{6}$$

and

$$\epsilon \sim MVN_{2n}(\boldsymbol{0}, \tau^2 \boldsymbol{I}) \tag{7}$$

where $\boldsymbol{A}$ has dimension $2n \times n^2$, $\boldsymbol{I}$ is the $2n \times 2n$ identity matrix, and ϵ (measurement error) is a $2n \times 1$ vector of iid normal random variables with variance τ^2. The conditional joint distribution of the row and column-pool measurements is then

$$\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} | \boldsymbol{Z} \sim MVN_{2n}(\mu(\boldsymbol{Z}), \boldsymbol{\Sigma}(\boldsymbol{Z})) \tag{8}$$

where

$$\mu(\boldsymbol{Z}) = \boldsymbol{A}\nu(\boldsymbol{Z}) \tag{9}$$

$$= \begin{pmatrix} \frac{1}{n}(\lambda \sum_{j=1}^{n} z_{1j} + \theta \sum_{j=1}^{n}(1 - z_{1j})) \\ \frac{1}{n}(\lambda \sum_{j=1}^{n} z_{2j} + \theta \sum_{j=1}^{n}(1 - z_{2j})) \\ \vdots \\ \frac{1}{n}(\lambda \sum_{i=1}^{n} z_{in} + \theta \sum_{i=1}^{n}(1 - z_{in})) \end{pmatrix} \tag{10}$$

$$= \begin{pmatrix} \frac{\lambda-\theta}{n} \sum_{j=1}^{n} z_{1j} + \theta \\ \frac{\lambda-\theta}{n} \sum_{j=1}^{n} z_{2j} + \theta \\ \vdots \\ \frac{\lambda-\theta}{n} \sum_{i=1}^{n} z_{in} + \theta \end{pmatrix} \tag{11}$$

$$= \frac{\lambda - \theta}{n} \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Z}^T \end{pmatrix} \mathbf{1}_n + \theta \mathbf{1}_{2n} \tag{12}$$

and

$$\boldsymbol{\Sigma}(\boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{\Gamma}(\boldsymbol{Z})\boldsymbol{A}^T + \tau^2 \boldsymbol{I} =$$

$$\frac{1}{n^2}\begin{pmatrix} \sum_{j=1}^{n}(\sigma_f^2)^{z_{1j}}(\sigma_n^2)^{1-z_{1j}} & \dots & 0 & (\sigma_f^2)^{z_{11}}(\sigma_n^2)^{1-z_{11}} & \dots & (\sigma_f^2)^{z_{1n}}(\sigma_n^2)^{1-z_{1n}} \\ 0 & \dots & 0 & (\sigma_f^2)^{z_{21}}(\sigma_n^2)^{1-z_{21}} & \dots & (\sigma_f^2)^{z_{2n}}(\sigma_n^2)^{1-z_{2n}} \\ \vdots & & \vdots & \vdots & & \vdots \\ 0 & \dots & \sum_{j=1}^{n}(\sigma_f^2)^{z_{nj}}(\sigma_n^2)^{1-z_{nj}} & (\sigma_f^2)^{z_{n1}}(\sigma_n^2)^{1-z_{n1}} & \dots & (\sigma_f^2)^{z_{nn}}(\sigma_n^2)^{1-z_{nn}} \\ (\sigma_f^2)^{z_{11}}(\sigma_n^2)^{1-z_{11}} & \dots & (\sigma_f^2)^{z_{n1}}(\sigma_n^2)^{1-z_{n1}} & \sum_{i=1}^{n}(\sigma_f^2)^{z_{i1}}(\sigma_n^2)^{1-z_{i1}} & \dots & 0 \\ (\sigma_f^2)^{z_{12}}(\sigma_n^2)^{1-z_{12}} & \dots & (\sigma_f^2)^{z_{n2}}(\sigma_n^2)^{1-z_{n2}} & 0 & \dots & 0 \\ \vdots & & \vdots & \vdots & & \vdots \\ (\sigma_f^2)^{z_{1n}}(\sigma_n^2)^{1-z_{1n}} & \dots & (\sigma_f^2)^{z_{nn}}(\sigma_n^2)^{1-z_{nn}} & 0 & \dots & \sum_{j=1}^{n}(\sigma_f^2)^{z_{in}}(\sigma_n^2)^{1-z_{in}} \end{pmatrix} + T \tag{13}$$

where $T = \tau^2 \boldsymbol{I}$. In situations where we have differential measurement error between normals and failures, we instead let

$$T = diag\left(\sum_{j=1}^{n}(\tau_f^2)^{z_{1j}}(\tau_n^2)^{1-z_{1j}}, \dots, \sum_{j=1}^{n}(\tau_f^2)^{z_{nj}}(\tau_n^2)^{1-z_{nj}}, \sum_{i=1}^{n}(\tau_f^2)^{z_{i1}}(\tau_n^2)^{1-z_{i1}}, \dots, \sum_{j=1}^{n}(\tau_f^2)^{z_{in}}(\tau_n^2)^{1-z_{in}}\right) \tag{14}$$

where τ_f^2 is the measurement error variance associated with failures and τ_n^2 is the measurement error variance associated with normals.

The complete data likelihood (the joint distribution of $\boldsymbol{r}$, $\boldsymbol{c}$, and $\boldsymbol{Z}$) can be written as the product of the conditional and marginal distributions as follows:

$$f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{Z}|\boldsymbol{\Theta}) = f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{Z}, \boldsymbol{\Theta}) f(\boldsymbol{Z}|\boldsymbol{\Theta}) \tag{15}$$

$$= (2\pi)^{-n}|\boldsymbol{\Sigma}|^{-\frac{1}{2}} exp(-\frac{1}{2}(\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} - \mu(\boldsymbol{Z}))^T \boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix} \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} - \mu(\boldsymbol{Z}))) \times p^{\Sigma z_{ij}}(1-p)^{\Sigma(1-z_{ij})} \tag{16}$$

where $\boldsymbol{\Theta} = (\lambda, \theta, \sigma_f, \sigma_n, \tau, p)$, and Σ represents the summation over both i and j.

3.3 Implementing the EM Algorithm

The EM algorithm begins by maximizing the expected complete-data log likelihood over the unknown parameters. In the context of matrix pooling the number of observed data points is small (on the order of $2n$ for n between 6 and 12) and hence insufficient to estimate all six distributional parameters in this model. We therefore assume that all parameters except p can be well estimated using pre-existing data and scientific knowledge of testing devices. This seems especially reasonable if we imagine that the lab is doing repeated testing from the same population and can learn about the parameters over time. The expected complete-data log likelihood, conditional on the observed data, is

$$\begin{aligned}
E[l(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{r},\boldsymbol{c})|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] &= E[log(f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&= E[-nlog(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))^T\boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z})) \\
&\quad +\Sigma z_{ij}log(p) + \Sigma(1-z_{ij})log(1-p)|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&= -nlog(2\pi) - \frac{1}{2}log|\boldsymbol{\Sigma}| \\
&\quad -\frac{1}{2}E[(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))^T\boldsymbol{\Sigma}(\boldsymbol{Z})^{-1}(\begin{pmatrix}\boldsymbol{r}\\ \boldsymbol{c}\end{pmatrix} - \mu(\boldsymbol{Z}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&\quad +log(p)\Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \\
&\quad +log(1-p)(n^2 - \Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]) \qquad (17)
\end{aligned}$$

Only the final two terms

$$log(p)\Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] + log(1-p)(n^2 - \Sigma E[z_{ij}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]) \qquad (18)$$

are associated with the unknown parameter of interest p, and so, the expected complete-data likelihood, conditional on $\boldsymbol{r}$, $\boldsymbol{c}$ and parameter estimates is a linear function of

$$E[\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{\Theta}]. \tag{19}$$

However, the marginal distribution of $\boldsymbol{r}$ and $\boldsymbol{c}$, $f(\boldsymbol{r}, \boldsymbol{c}|\boldsymbol{\Theta})$ is not known and can only be computed by summing over all 2^{n^2} values of $\boldsymbol{Z}$, and hence it would be difficult to compute $E[\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{\Theta}]$ analytically. Instead we estimate the expectation by summing over only the configurations of $\boldsymbol{Z}$ which could potentially yield the observed row and column values. All other configurations would have very low probability, and hence would contribute very little to the expectation equation.

3.4 Approximate expectation of $\boldsymbol{Z}$ (E Step)

For each row and column we first estimate the number of failures based on measured concentrations r_i and c_j and on our preliminary estimates $\hat{\lambda}$ and $\hat{\theta}$ of λ and θ. Let f_{r_i} be the approximate number of failures that would be expected to yield the observed value of r_i. Choose f_{r_i} so that

$$1/n(\hat{\lambda}(f_{r_i} - 1) + \hat{\theta}(n - f_{r_i} + 1) \leq r_i \leq 1/n(\hat{\lambda} f_{r_i} + \hat{\theta}(n - f_{r_i}), \tag{20}$$

unless

$$r_i \leq 1/n(\hat{\lambda} + \hat{\theta}(n - 1)) \tag{21}$$

in which case $f_{r_i} = 1$ if $r_i > (t^*/n)$ and $f_{r_i} = 0$ otherwise, where t^* is the failure threshold. For each column j, estimate the number of failures f_{c_j} the same way. This method will tend to overestimate the numbers of failures in each row and column, and, ignoring measurement error, will prevent any failures from being missed. Let $\{\boldsymbol{Z}'\}$ be the set of all $n \times n$ matrix configurations of failures and non-failures (1's and 0's) such that the number of failures in each row and column is less than or equal to f_{r_i} and f_{c_j}. Then compute (19) as

$$E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] \quad = \quad \sum_{\{\boldsymbol{Z}'\}} \boldsymbol{Z}' f(\boldsymbol{Z}'|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}) \tag{22}$$

where

$$f(\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}) \approx f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}) / \sum_{\{\boldsymbol{Z}'\}} f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}'|\boldsymbol{\Theta}) \tag{23}$$

3.5 The M-Step

At each iteration of the E-M algorithm we then maximize the expected complete-data log likelihood over p. Differentiaing (17) by p, and defining $\hat{z}_{ij} = E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}]_{ij}$, we have

$$\frac{\partial}{\partial p} E[l(\boldsymbol{\Theta}|\boldsymbol{Z},\boldsymbol{r},\boldsymbol{c}])|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}] = \frac{1}{p}\Sigma\hat{z}_{ij} - \frac{1}{1-p}(n^2 - \Sigma\hat{z}_{ij}) \tag{24}$$

which, after setting equal to 0 and solving for p gives

$$\hat{p} = \frac{1}{n^2}\Sigma\hat{z}_{ij}. \tag{25}$$

EM estimation proceeds as usual by choosing starting values $\boldsymbol{\Theta}_{(0)} = (\hat{\lambda}, \hat{\theta}, \hat{\sigma}, \hat{\tau}, p_{(0)})$, and estimating $E[log(f(\boldsymbol{r},\boldsymbol{c},\boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r},\boldsymbol{c},\boldsymbol{\Theta}_{(0)}]$ by using the method described in Section 3.4. Then maximize over the unknown portion of the parameter space to get $\hat{\boldsymbol{\Theta}}_{(1)}$, re-estimate the expected log likelihood by estimating $E[\boldsymbol{Z}|\boldsymbol{r},\boldsymbol{c},\hat{\boldsymbol{\Theta}}_{(1)}]$, and so on, iterating until the parameter estimates converge. "Convergence" is defined as the iteration where $\hat{p}_{(n)}$ changes by less than 0.0001 from one iteration to the next.

3.6 Test individual specimens, and repeat EM

Once the EM algorithm has converged, identify the failure configuration $\boldsymbol{Z}'$ with the highest conditional probability, and test each individual specimen indicated in $\boldsymbol{Z}'$ by a 1. Let $\boldsymbol{y}^*$ be the $k \times 1$ vector of measurements for all individually tested specimens, and let $\boldsymbol{z}^*$ be the corresponding elements of $\boldsymbol{Z}$. Now repeat the EM algorithm by finding

$$E[log(f(\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{Z}|\boldsymbol{\Theta}))|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}] \tag{26}$$

which again reduces to finding

$$E[\boldsymbol{Z}|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}_{(0)}] = \sum_{\{\boldsymbol{Z}'\}} \boldsymbol{Z}' f(\boldsymbol{Z}'|\boldsymbol{r}, \boldsymbol{c}, \boldsymbol{y}^*, \boldsymbol{\Theta}_{(0)}) \tag{27}$$

$$\tag{28}$$

where $\boldsymbol{\Theta}_{(0)}$ is set equal to the final $\boldsymbol{\Theta}_{(n)}$ from the prior round of EM. From (3.2) we can derive that

$$\begin{pmatrix} \boldsymbol{y}^* \\ \boldsymbol{r} \\ \boldsymbol{c} \end{pmatrix} |\boldsymbol{Z} \sim MVN_{2n+k}(\mu(\boldsymbol{Z})^*, \boldsymbol{\Sigma}(\boldsymbol{Z})^*) \tag{29}$$

where

$$\mu(\boldsymbol{Z})^* = \begin{pmatrix} \lambda \boldsymbol{z}^* + \theta(\mathbf{1}_k - \boldsymbol{z}^*) \\ \frac{\lambda - \theta}{n} \begin{pmatrix} \boldsymbol{Z} \\ \boldsymbol{Z}^T \end{pmatrix} \mathbf{1}_n + \theta \mathbf{1}_{2n} \end{pmatrix} \tag{30}$$

and $\boldsymbol{\Sigma}(\boldsymbol{Z})^*$ is straightforward expansion of (13). The set $\{\boldsymbol{Z}'\}$ now depends on the observed values $\boldsymbol{y}^*$. Define r_i' as the approximate concentration in the ith row pool if the target substance in the tested cells were removed:

$$r_i' = r_i - \frac{1}{n} \sum_{j=1}^{n} y_{ij}^*. \tag{31}$$

For the ith row, estimate the number of failures in untested cells by choosing f_{r_i}' such that

$$1/n(\hat{\lambda}(f_{r_i}' - 1) + \hat{\theta}(n - f_{r_1}' + 1) \leq r_i' \leq 1/n(\hat{\lambda} f_{r_i}' + \hat{\theta}(n - f_{r_1}'), \tag{32}$$

unless

$$r'_i \leq 1/n(\hat{\lambda} + \hat{\theta}(n-1)) \tag{33}$$

in which case $f'_{r_i} = 1$ if $r'_i > (t^*/n)$ and $f'_{r_i} = 0$ otherwise. Now the estimated number of failures in row i is

$$f_{r_i} = \sum_{j=1}^{n} 1_{[y^*_{ij} > t^*]} + f'_{r_i}, \tag{34}$$

and the updated number of failures in each column is compute similarly.

Upon convergence of each subsequent round of the EM algorithm, identify the most probable configuration $\boldsymbol{Z}'$ and test any indicated specimens that have not previously been tested. If there are no further candidate specimens to test, the process is complete.

4 Simulation studies

In order to determine whether this methodology can reduce turn-around time and improve efficiency as compared to previously studied pooling methods, we simulate random samples of n^2 specimens for a variety of different values of n and prevalence p. We compare the EM method to the Simple Search method (May et al., 2010) and to a modified version of the Simple Search, comparing efficiency, turn-around time, Sensitivity, and Negative Predictive Value (NPV). "Efficiency" is defined as the percent reduction in the number of tests needed to detect all failures, as compared to testing each individual separately.

4.1 Comparable Methods

The "Simple Search" method developed by May et al. (2010) involves testing the row and column pools of a $n \times n$ matrix array and then testing the individual cell at the intersection of the row and column with the highest combined concentration that exceeds the failure threshold t divided by n. The tested value is then subtracted from the corresponding row and column pool values, and the cell with the next

highest combined row and column concentration is tested. This process continues, testing one cell at a time, until there are no cells having both a row and column that exceed t/n.

The "Modified Simple Search" is a variation on the Simple Search method designed to reduce turn-around time. Whereas the Simple Search method identifies and tests the single cell with the highest combined row and column value, the Modified Simple Search further identifies and tests additional cells in unique rows and columns. Once the first cell is selected, the cell with the next highest combined row and column value is chosen from the remaining cells in distinct rows and columns. This process is repeated until approximately $n/2$ cells are identified for testing, and all $n/2$ cells are tested at once. Because more individual cells are identified for testing at each round, fewer testing rounds are necessary.

4.2 Normally distributed data

For the first set of simulation studies, target values were generated as a mixture of normal random variables with distinct means and variances, and with a common, small, measurement error variance. Two hundred simulated datasets were generated for each of a variety of fixed prevalence levels ranging from 0.01 to 0.16. Target values were simulated according to model (3.2), with $\lambda = 3000$, $\theta = 200$, $\sigma_f = 200$, $\sigma_n = 50$, and $\tau = 5$. The modelling procedure used the following parameter estimates: $\hat{\lambda} = 3100$ $\hat{\sigma}_f = 210$, $\hat{\theta} = 220$, $\hat{\sigma}_n = 48$ and $\hat{\tau} = 5$, with failure threshold $t^* = 1000$. Efficiency, turn-around time and sensitivity results are shown in Table 1 and Appendix A.

Table 1: Simulation results for 8x8 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
Rounds - EM	1.0	1.3	1.4	1.7	1.9	2.1	2.4	2.4	2.5	2.6
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.1	2.5	3.2	3.9	4.4	5.0	5.7
Rounds - Simple Search	1.0	2.4	3.7	5.7	7.7	10.0	12.6	14.3	16.8	19.2
Efficiency - EM	73.4	71.1	68.9	66.1	62.9	59.8	56.4	54.1	50.8	47.4
Efficiency - Mod. Simple Search	73.4	69.3	67.7	64.3	61.2	57.1	52.6	49.4	45.8	41.6
Efficiency - Simple Search	73.4	71.3	69.2	66.2	63.0	59.4	55.3	52.6	48.8	44.9
Sensitivity - EM	100	100	100	100	100	100	99.7	100	100	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Individual Testing	100	100	100	100	100	100	99.9	100	100	100

For prevalences less than about 6% all three methods perform similarly well in terms of efficiency. For prevalences above 10% the EM outperforms both Simple Search methods, with the advantage of EM increasing for higher prevalences. At the highest prevalence level tested (15.6%, or 10 failures in an 8×8 matrix), the EM method achieved 47.4% efficiency, while the SS and MSS methods were only 44.9% and 41.6% efficient, respectively.

In terms of turn-around time the EM method outperforms Simple Search methods even at low prevalances, and the benefit increases dramatically with increasing prevalence. At 15.6% prevalence the EM cuts the turn-around time in half as compared to the MSS, and improves on the SS method by 86%.

4.3 Skewed Data - HIV-1 Viral Load Example

In resource-wealthy settings, HIV-infected individuals who are being treated with anti-retroviral medications (ART) are routinely monitored for virologic failure, defined as a detectable proliferation of HIV virus in the blood despite treatment. Individuals experiencing virologic failure may decline into worse health and experience AIDS, and they also may present an increased risk of transmitting treatment-resistant virus to sexual partners. In the context of low-income countries, regular viral-load testing (performed by reverse transcriptase polymerase chain reaction (RT PCR)) is expensive and time consuming. Limited HIV funds must also be allocated to identifying HIV cases and providing treatment, education, prevention interventions, and a host of other services. It is critical therefore in these settings to minimize the cost of viral monitoring.

In general we do not expect true viral-load values to follow a normal distribution. Empirical data suggest that viral-load values tend to be skewed and are often well described by the lognormal distribution. In addition, assay measurement error tends to be constant on the log scale, with assay standard deviations of approximately 0.12 on the log_{10} scale (Brambilla et al., 1999), suggesting substantial variation for large viral loads. To test robustness to distributional assumptions, we also ran a simulation study where viral load values are not assumed to follow a normal mixture.

Simulated values for individuals with supressed viral load values were generated as exponential random variables with a mean of 50. Failure values were generated as lognormal with mean 3.2 and standard

deviation 0.25 (on the log_{10} scale), and shifted by +1000, giving an actual mean of about 2,585. Measurement error was generated on the log_{10} scale as normal with mean 0 and standard error 0.12. Because measurement error is applied on the log_{10} scale, we use the second measurement-error variance term T as specified in (14), which accommodates the much larger measurement error among failures on the standard scale. When running the EM modeling procedure, the following parameter estimates were assumed: $\hat{\lambda} = 2800$, $\hat{\sigma}_f = 750$, $\hat{\tau}_f = 900$, $\hat{\theta} = 50$, $\hat{\sigma}_n = 50$ and $\hat{\tau}_n = 5$, with failure threshold $t^* = 1000$.

Two hundred datasets were generated for each prevalence level, with prevalence levels fixed so that each dataset contained the same, known, number of failures. Simulation results are displayed in Table 2. Although turn-around times and efficiency look favorable for the EM method, these promising values come at the expense of a dramatic loss in sensitivity. All three methods show reduced sensitivity as compared with the normal-data simulations, however sensitivity for the the EM method is particularly low, at only 74% at the 16% prevalence level.

Table 2: Simulation results for 8x8 matrix pools. Skewed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents) (Note - these results are not final. Simulations for higher prevalence levels were not completed (stopped due to futility). n for top three prevalence levels are 197, 84, 4.)

	Prevalence (%)									
	1.6	3.1	4.7	6.2	7.8	9.4	10.9	12.5	14.1	15.6
Rounds - EM	0.9	1.2	1.5	2.0	2.4	3.0	3.2	3.7	4.0	4.0
Rounds - Mod. Simple Search	1.0	1.0	1.6	2.2	2.8	3.5	4.0	4.7	5.3	6.0
Rounds - Simple Search	1.1	2.4	4.0	5.9	7.8	10.2	12.3	14.5	17.0	19.2
Efficiency - EM	73.5	71.2	68.5	65.3	62.2	58.5	56.0	52.4	50.3	50.0
Efficiency - Mod. Simple Search	73.1	69.4	66.8	63.2	59.5	54.9	51.6	47.5	43.9	39.5
Efficiency - Simple Search	73.3	71.3	68.7	65.8	62.7	59.0	55.8	52.3	48.4	44.9
Sensitivity - EM	89.5	90.5	90.2	87.8	87.2	84.8	80.0	77.5	74.7	62.5
Sensitivity - Mod. Simple Search	98.5	99.5	96.5	96.2	94.4	93.7	91.7	90.8	89.3	95.0
Sensitivity - Simple Search	98.5	96.5	94.5	94.2	93.6	92.2	90.9	90.2	89.0	92.5
Sensitivity - Individual Testing	99.5	100	98.7	98.9	99.7	99.2	99.0	99.3	99.2	100

5 Discussion

These results suggest that for biological samples with normally distributed assay values, the EM-Matrix Pooling approach is quite effective. Although the cost savings associated with this approach is not substantial (about 3% at high prevalence levels), the time savings associated with shorter turn-around times

is large (about 85% compared to Simple Search and 50% compared to Modified Simple Search). In cases where observed assay data is likely skewed such as HIV-1 viral load concentrations which tend to be lognormal, and particularly where measurement error is quite large, this method breaks down and does not yield high enough sensitivity to warrant use. The problem seems to be that the testing algorithm finishes too quickly, and is unable to recognize that additional, untested failures are present in the matrix. Perhaps it should not be surprising that large measurement error associated with pooled data would make inference about individual samples very difficult.

We did run simulations where the assay data was assumed to be lognormal, but the measurement error was small, and the the EM method performed well in terms of sensitivity and turn-around time. However, efficiency was not improved as compared to the Simple Search methods. (Results not shown.) Still, in situations where each round of testing requires a substantial amount of time, this method could be quite useful.

One limitation to our method is the requirement that most of the distributional parameters are known or well estimated. Although it would be nice to estimate or update these parameters based on pooled data, it is unlikely that the very sparse data associated with matrix-pooled measurements would provide good information. However, in a laboratory setting where large numbers of biological samples are processed each day, it would be quite reasonable to estimate distributional parameters for populations that are sampled regularly over time. In addition, it would not be difficult to obtain repeated tests on certain samples in order to estimate measurement-error distributions. Assay calibration may require this.

Further investigation with normally distributed assay data should include an assessment of model performance under varying degrees of overlap between the failure distribution and non-failure distribution. The simulated examples reported here use distributions that have little overlap. An important limiting factor will be the need for the failure mean λ divided by n to be larger than the non-failure mean θ. If this is not the case, then a row with a single failure and the rest undetectable would be indistinguishable from a row of all non-failures with average concentrations. In general, it will likely be the case that for a given λ and θ, only pooling matrices of size n^* or less would be feasible, where n^* is the largest n for which $\lambda/n > c\theta$ for some constant c. A reasonable value of c would need to be investigated. Sensitivity to larger measurement error should also be explored.

In conclusion, this method was developed in the hope that HIV-1 viral load monitoring could be made faster and more efficient. Although the method works for normal data with minimal measurement error, unfortunately it does not seem to work well for skewed data with substantial measurement error, such as HIV-1 viral loads.

Appendix A - Simulation results for varying matrix sizes.

Table 3: Simulation results for 5x5 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	4	8	12	16	20	24	28	32	36	40
Rounds - EM	1.0	1.2	1.4	1.6	1.7	1.8	1.9	1.9	2.0	2.1
Rounds - Mod. Simple Search	1.0	1.2	1.9	2.3	2.9	3.5	4.1	4.6	5.1	5.5
Rounds - Simple Search	1.0	2.3	3.8	5.4	7.1	8.8	10.2	11.9	13.5	15.0
Efficiency - EM	56.0	50.4	44.9	38.8	32.9	26.9	21.8	17.2	10.3	5.1
Efficiency - Mod. Simple Search	56.0	48.6	42.5	35.1	28.6	21.4	14.5	8.4	2.7	-1.6
Efficiency - Simple Search	56.0	50.7	44.9	38.3	31.7	24.7	19.1	12.6	5.9	-0.0
Sensitivity - EM	100	100	100	99.9	100	100	100	100	99.8	100
Sensitivity - Mod. Simple Search	100	100	100	99.9	100	100	100	100	99.9	100
Sensitivity - Simple Search	100	100	100	99.9	100	100	100	100	99.9	100
Sensitivity - Individual Testing	100	100	100	99.9	100	100	100	100	99.9	100

Table 4: Simulation results for 6x6 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	2.8	5.6	8.3	11.1	13.9	16.7	19.4	22.2	25	27.8
Rounds - EM	1.0	1.1	1.4	1.6	1.8	1.9	2.0	2.2	2.2	2.3
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.0	2.5	3.0	3.6	4.1	4.5	4.9
Rounds - Simple Search	1.0	2.2	3.9	5.5	7.6	9.5	11.1	13.2	14.9	16.4
Efficiency - EM	63.9	60.3	55.9	51.4	46.6	43.2	38.1	32.7	29.0	25.2
Efficiency - Mod. Simple Search	63.9	57.3	53.6	48.4	42.3	36.7	30.5	24.7	20.5	16.3
Efficiency - Simple Search	63.9	60.4	55.9	51.4	45.7	40.4	35.8	30.1	25.3	21.2
Sensitivity - EM	100	100	100	99.5	100	99.8	99.7	100	99.6	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	99.9	99.9	100	99.9	100
Sensitivity - Simple Search	100	100	100	100	100	99.9	99.9	100	99.9	100
Sensitivity - Individual Testing	100	100	100	100	100	99.9	99.9	100	99.9	100

Table 5: Simulation results for 7x7 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	2	4.1	6.1	8.2	10.2	12.2	14.3	16.3	18.4	20.4
Rounds - EM	1.0	1.2	1.4	1.6	1.9	2.0	2.2	2.4	2.3	2.5
Rounds - Mod. Simple Search	1.0	1.0	1.5	2.0	2.5	3.1	3.7	4.3	4.8	5.3
Rounds - Simple Search	1.0	2.3	3.8	5.7	7.9	9.7	12.1	13.9	16.0	18.1
Efficiency - EM	69.4	66.6	63.5	60.2	56.0	52.7	49.1	44.8	42.1	38.2
Efficiency - Mod. Simple Search	69.4	64.4	62.0	57.8	53.1	48.6	43.5	38.7	35.0	30.9
Efficiency - Simple Search	69.4	66.7	63.7	59.8	55.3	51.7	46.7	43.0	38.8	34.5
Sensitivity - EM	100	100	99.8	100	99.4	99.6	100	100	99.9	100
Sensitivity - Mod. Simple Search	100	100	99.8	100	99.8	99.9	100	100	99.9	100
Sensitivity - Simple Search	100	100	99.8	100	99.8	99.9	100	100	99.9	100
Sensitivity - Individual Testing	100	100	99.8	100	99.8	99.9	100	100	99.9	100

Table 6: Simulation results for 9x9 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1.2	2.5	3.7	4.9	6.2	7.4	8.6	9.9	11.1	12.3
Rounds - EM	1.0	1.2	1.5	1.8	2.0	2.2	2.4	2.6	2.8	2.9
Rounds - Mod. Simple Search	1.0	1.0	1.4	2.0	2.3	2.9	3.4	3.9	4.4	5.0
Rounds - Simple Search	1.0	2.4	3.9	6.0	7.9	10.4	13.1	15.3	17.4	20.1
Efficiency - EM	76.5	74.7	72.7	70.4	68.1	65.5	62.8	59.8	56.7	54.5
Efficiency - Mod. Simple Search	76.5	73.3	71.0	68.4	65.9	62.3	59.1	55.8	53.5	49.3
Efficiency - Simple Search	76.5	74.9	73.0	70.4	68.0	64.9	61.6	58.9	56.2	52.9
Sensitivity - EM	100	100	100	99.5	100	100	99.6	100	100	100
Sensitivity - Mod. Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Simple Search	100	100	100	100	100	100	99.9	100	100	100
Sensitivity - Individual Testing	100	100	100	100	100	100	99.9	100	100	100

Table 7: Simulation results for 10x10 matrix pools. Normally distributed data, with 200 simulated datasets per prevalence level. (Efficiency and Sensitivity are percents)

	Prevalence (%)									
	1	2	3	4	5	6	7	8	9	10
Rounds - EM	1.1	1.2	1.5	1.9	2.1	2.4	2.6	2.7	2.9	3.0
Rounds - Mod. Simple Search	1.0	1.0	1.3	1.8	2.2	2.6	3.1	3.5	4.0	4.5
Rounds - Simple Search	1.1	2.5	4.1	6.1	8.2	10.5	13.2	15.8	18.3	20.3
Efficiency - EM	78.9	77.6	75.9	74.0	72.0	69.8	66.9	65.1	62.6	60.3
Efficiency - Mod. Simple Search	78.3	75.6	73.5	71.8	69.2	66.7	63.8	61.5	57.9	55.3
Efficiency - Simple Search	78.9	77.5	75.9	73.9	71.8	69.5	66.8	64.2	61.7	59.7
Sensitivity - EM	100	90.0	90.0	94.6	95.2	97.8	99.6	100	100	100
Sensitivity - Mod. Simple Search	100	100	99.8	100	99.9	100	100	100	100	100
Sensitivity - Simple Search	100	100	99.8	100	99.9	100	100	100	100	100
Sensitivity - Individual Testing	100	100	99.8	100	99.9	100	100	100	100	100

References

Donald Brambilla, Patricia Reichelderfer, James Bremer, David Shapiro, Ronald Hershow, David Katzenstein, Scott Hammer, Brooks Jackson, and Ann Collier. The contribution of assay variation and biological variation to the total variability of plasma hiv-1 rna measurements. *AIDS*, 1999.

Robert Dorfman. The detection of defective members of large populations. *Annals of Mathematical Statistics*, 1943.

Susanne May, Anthony Ganst, Richard Haubrich, Constance Benson, and Davey Smith. Pooled nucleic acid testing to identify anteretroviral treatment failure during hiv infevtion. *Journal of Acquired Immune Defficiency Syndrome*, 2010.

RM Phadarfod and Aidan Sudbury. The use of a square array scheme in blood testing. *Statistics in Medicine*, 1994.