University of South Carolina

Hierarchical group testing for multiple infections

Peijie Hou and Joshua M. Tebbs

Background

Group testing, where individuals are initially tested in pools, is often used to screen a large number of individuals for infectious diseases.

Triggered by the development of assays that detect multiple infections, large-scale screening programs now involve testing individuals in pools for multiple infections simultaneously. Tebbs et al. (2013) have recently evaluated the performance of a two-stage hierarchical algorithm that is used to screen for chlamydia and gonorrhea as part of the Infertility Prevention Project (IPP) in the US.

Our interest is to generalize this work (two infections) to accommodate a larger number of stages.

1. We use Markov chain framework to derive operating characteristics.
2. We use EM algorithm to estimate the population prevalence.

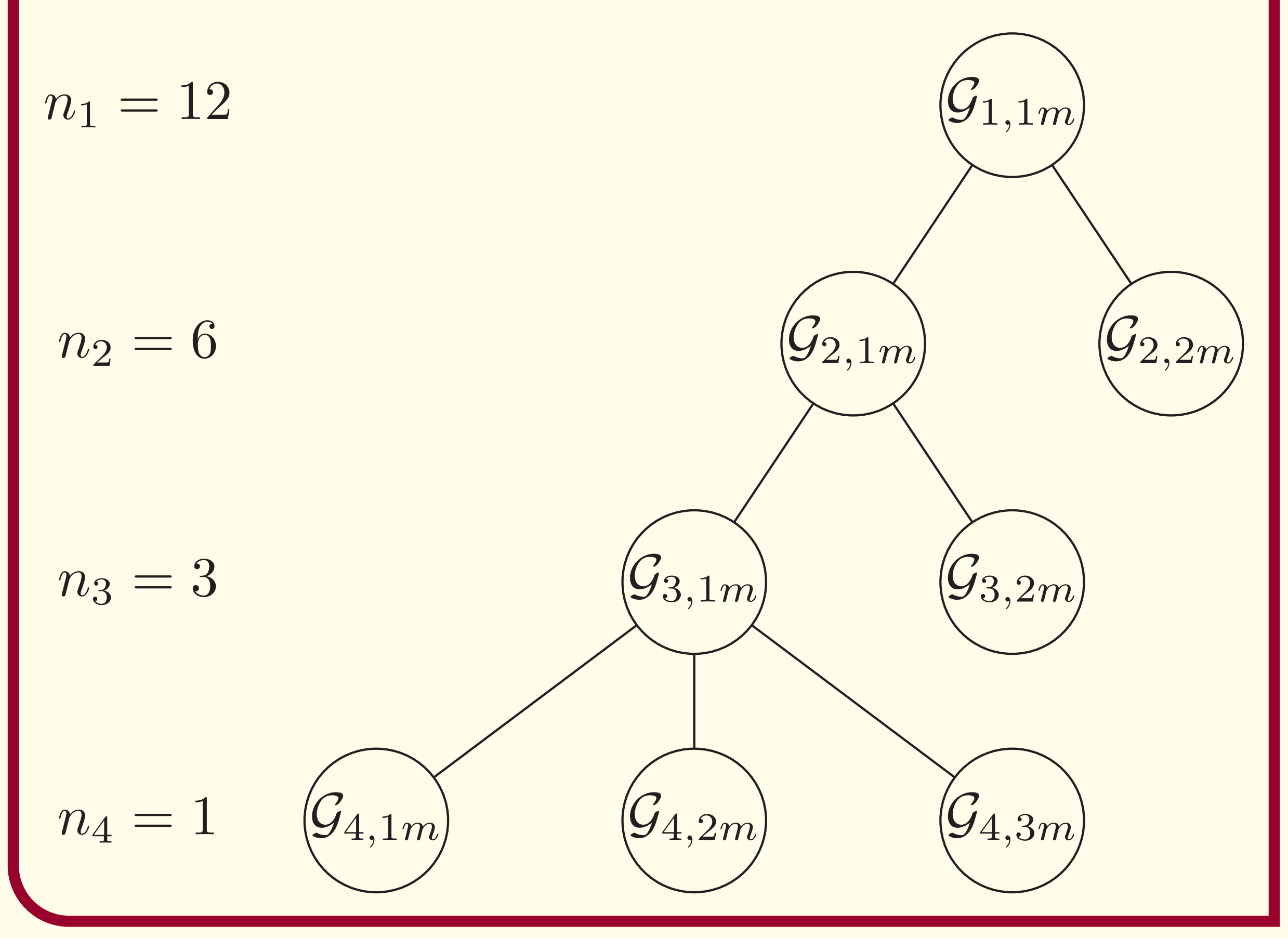
Contribution

1. We offer compelling evidence that higher-stage algorithm can provide large cost savings, especially when disease probabilities are low.

2. Higher-stage algorithm can provide prevalence estimates that are as efficient as those from the two-stage algorithm in Tebbs et al. (2013).

Testing algorithm

An example of four-stage algorithm with master pool size $n_1 = 12$, $\mathcal{G}_{s,im}$ denotes the ith pool at the sth stage of mth master pool.



Notation and assumptions

Notation:

1. n_s: pool size at sth stage, for $s = 1, \ldots, S$.
2. $\tilde{Y}_{ljm} = 1$: true status of jth infection for lth individual in mth master pool, where $l = 1, \ldots, n_1$, $m = 1, \ldots, M$ and $j = 1, 2$.

Assumptions:

1. All individual specimens are i.i.d.
2. Sensitivity ($S_{e:j}$) and specificity ($S_{p:j}$) of the assay are known and do not depend on the size of the pool being tested.
3. If the true statuses are given, the test responses are mutually independent.

Expected number of tests (efficiency)

The true status of the master pool transits to the true status of $\mathcal{G}_{s,im}$ is a finite space time-inhomogeneous Markov chain, with *states* $\Omega = \{00, 01, 10, 11\}$ and each stage is considered to be a *step*. We derive the expected number of tests needed to decode all individuals in mth master pool with an $S-$stage algorithm:

$$E\left(T_m^{(S)}\right) = 1 + \sum_{s=2}^{S} (n_1/n_s) \left\{ J_4' \times \mathcal{M} \times \mathcal{P} \times \left[\prod_{t=0}^{s-2} \left(\boldsymbol{\pi}^{(t)} \mathcal{P} \right) \right] \times J_4 \right\}$$

where $\mathcal{M}$ denotes the probability distribution of the starting state (true status of the master pool), $\mathcal{P}$ is the classification probability matrix that a pool with certain true status being tested positive for at least one infection, $J_4 = (1,1,1,1)'$, $\boldsymbol{\pi}^{(0)} = \mathcal{P}^{-1}$ and the stage-dependent transition matrix $\boldsymbol{\pi}^{(s)}$ has the form below, where column label denotes the true status of the parent pool and the row labels denote the true status of the sub-pool:

$$\boldsymbol{\pi}^{(s)} = \begin{array}{c} \\ (1,1) \\ (1,0) \\ (0,1) \\ (0,0) \end{array} \begin{array}{c} \begin{array}{cccc} (1,1) & (1,0) & (0,1) & (0,0) \end{array} \\ \begin{pmatrix} \pi_{11}^{(s)} & \pi_{12}^{(s)} & \pi_{13}^{(s)} & \pi_{14}^{(s)} \\ 0 & \pi_{22}^{(s)} & 0 & \pi_{24}^{(s)} \\ 0 & 0 & \pi_{33}^{(s)} & \pi_{34}^{(s)} \\ 0 & 0 & 0 & 1 \end{pmatrix} \end{array}$$

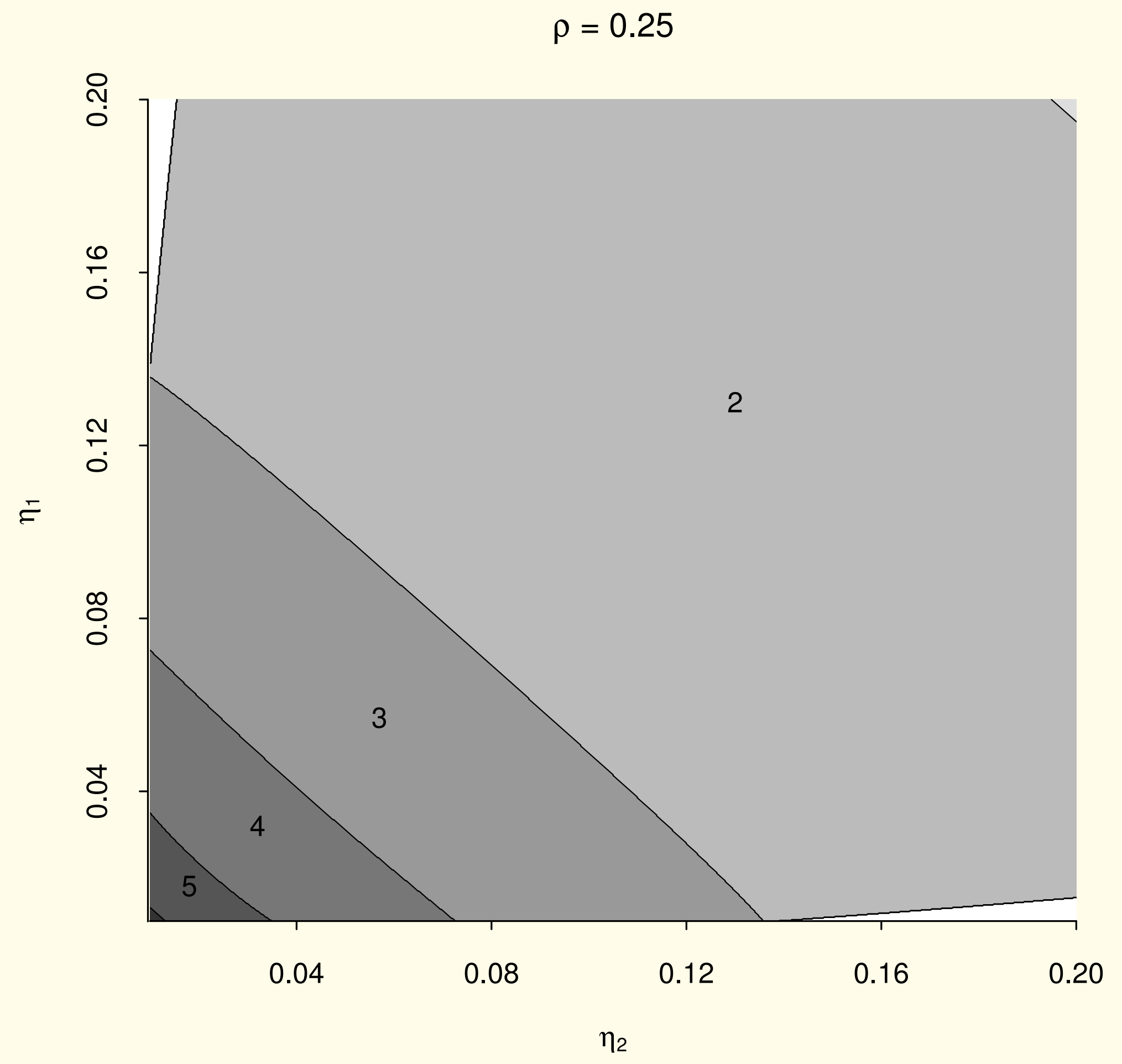
Remark: all probabilities in above matrix are derived in closed-form.

Optimal grouping strategy

With a closed-form expression available for efficiency, we are able to choose the optimal number of stages S and optimal master pool size n_1^* for a specific pool-splitting strategy (halving).

- For a given stage S, choose the optimal master pool size n_1^* such that $n_1^{-1} E(T_m^{(S)})$ is minimized (expected number of tests *per-individual*).
- Choose the number of stage S that produces the smallest expected number of tests per-individual.

Opt-stage determination



The optimal number of stage for different values of η_1 and η_2, where the correlation $\rho = \mathrm{corr}(\tilde{Y}_{l1m}, \tilde{Y}_{l1m})$ is set to be 0.25, $\eta_1 = P(\widetilde{Y}_{l1m} = 1) = p_{10} + p_{11}$ and $\eta_2 = P(\widetilde{Y}_{l2m} = 1) = p_{01} + p_{11}$. Values of $(\eta_1, \eta_2)'$ in the white regions are not possible.

Classification accuracy

Pooling sensitivity $PS_{e:j}$ (pooling specificity $PS_{p:j}$) for the jth infection is the probability an individual is classified as positive (negative) for the jth infection given that the individual is truly positive (negative). They are derived in the same manner of efficiency using Markov chain framework.

Estimation

We generalize the MCEM algorithm proposed by Tebbs et al. (2013) to estimate the prevalence $\boldsymbol{\vartheta} = (p_{00}, p_{10}, p_{01})'$ by treating the individual true statuses $\tilde{Y}_{ljm}$ as "missing data". The variance-covariance matrix estimate are obtained using the missing information principle and Louis's method (1982).

Real data: classification

To illustrate the potential benefits of the hierarchical algorithm, we use the 16440 swab specimens collected by Nebraska as part of the IPP in 2009. The optimal grouping strategy are determined by the "training data" in 2008.

		Male		Female	
		C	G	C	G
N		1910		14530	
$\bar{T}(n_1^*)$	2-stage	**1608.8 (3)**		7709.6 (4)	
	3-stage	1715.8 (6)		7120.4 (6)	
	4-stage	1827.8 (12)		**7091.3 (12)**	
PS_e	2-stage	0.931	0.986	0.894	0.988
	3-stage	0.908	0.984	0.849	0.985
	4-stage	0.895	0.982	0.812	0.982
PS_p	2-stage	0.990	0.990	0.994	0.996
	3-stage	0.990	0.990	0.997	0.998
	4-stage	0.990	0.990	0.997	0.998

Real data: estimation

		2-stage		3-stage		4-stage	
Stratum	Prevalence	Estimate	SE	Estimate	SE	Estimate	SE
Male Swab ($N = 1910$)	p_{00}	0.791	0.0099	0.791	0.0099	0.790	0.0101
	p_{10}	0.140	0.0085	0.139	0.0085	0.139	0.0088
	p_{01}	0.052	0.0053	0.052	0.0053	0.053	0.0053
	p_{11}	0.017	0.0032	0.017	0.0031	0.017	0.0031
Female Swab ($N = 14530$)	p_{00}	0.924	0.0024	0.924	0.0024	0.924	0.0024
	p_{10}	0.063	0.0022	0.063	0.0022	0.063	0.0022
	p_{01}	0.007	0.0007	0.007	0.0007	0.007	0.0007
	p_{11}	0.006	0.0007	0.006	0.0006	0.006	0.0006