BIOMETRICS 000, 000–000
000 0000

Adaptive elastic net for group testing

Karl B. Gregory^{1,*}, Dewei Wang^{1,**}, and Christopher S. McMahan^{2,***}

[1]Department of Statistics, University of South Carolina, Columbia, SC 29208, U.S.A.

[2]Department of Mathematical Sciences, Clemson University, Clemson, SC 29634, U.S.A.

**email:* gregorkb@stat.sc.edu

***email:* deweiwang@stat.sc.edu

****email:* mcmaha2@clemson.edu

SUMMARY: For disease screening, group (pooled) testing can be a cost-saving alternative to one-at-a-time testing, with savings realized through assaying pooled biospecimen (e.g. urine, blood, saliva). In many group testing settings, practitioners are faced with the task of conducting disease surveillance. That is, it is often of interest to relate individuals' true disease statuses to covariate information via binary regression. Several authors have developed regression methods for group testing data, which is challenging due to the effects of imperfect testing. That is, all testing outcomes (on pools and individuals) are subject to misclassification, and individuals' true statuses are never observed. To further complicate matters, individuals may be involved in several testing outcomes. For analyzing such data, we provide a novel regression methodology which generalizes and extends the aforementioned regression techniques and which incorporates regularization. Specifically, for model fitting and variable selection, we propose an adaptive elastic net estimator under the logistic regression model which can be used to analyze data from any group testing strategy. We provide an efficient algorithm for computing the estimator along with guidance on tuning parameter selection. Moreover, we establish the asymptotic properties of the proposed estimator and show that it possesses "oracle" properties. We evaluate the performance of the estimator through Monte Carlo studies and illustrate the methodology on a chlamydia data set from the State Hygienic Laboratory in Iowa City.

KEY WORDS: Adaptive elastic net; Group testing; Model selection.

This paper has been submitted for consideration for publication in *Biometrics*

1. Introduction

Group testing is becoming a popular cost-saving alternative to individual-level testing in applications from infectious disease testing (Lewis et al., 2012; Krajden et al., 2014) to environmental monitoring (Heffernan et al., 2014). For example, the State Hygienic Laboratory (SHL) in Iowa City saved approximately $3.1 million during 2009–2014 after adopting group testing to screen female subjects for chlamydia (Jeffrey Benfer, SHL, personal communication). In general, group testing involves combining specimen collected from individuals into non-overlapping groups (master pools) for testing. Individuals belonging to pools which test negatively are diagnosed as negative at the expense of a single assay, while positive pools are resolved through further testing. For example, Dorfman (1943) suggested that positive pools be resolved through individual-level testing. If the binary characteristic of interest, e.g. infection status, is rare, group testing can result in substantial cost savings.

Statistics research in group testing has generally focused on developing either classification or estimation methods, with the latter being our interest (for a review of classification algorithms see Kim et al., 2007). Estimation based on group testing data traces back to Thompson (1962), who focused on estimating a population-level proportion. This particular problem has gained considerable interest, both historically and recently; see Liu et al. (2011). Extending earlier works, several authors have proposed various parametric (Farrington, 1992; Vansteelandt et al., 2000; Huang, 2009; Chen et al., 2009; McMahan et al., 2012) and nonparametric (Delaigle and Meister, 2011; Delaigle et al., 2014; Wang et al., 2014; Delaigle and Hall, 2015) regression methodologies for group testing data. A drawback to the aforementioned methodologies is that they were designed only for analyzing test results obtained from assaying non-overlapping master pools; i.e., they cannot incorporate data from resolving positive master pools, testing procedures with overlapping pools, or data from quality control testing (Gastwirth and Johnson, 1994; Kim et al., 2007; Krajden et al., 2014).

Developing general regression methods for group testing data is challenging, since individuals' true disease statuses are never observed; and the observed data consists of assay outcomes which are liable to error. Moreover, the complexity of the problem increases when individuals are involved in multiple testing outcomes, which occurs during retesting, quality control steps, and through implementing certain group testing protocols. Several authors have proposed regression methods which can incorporate these more complex data structures; e.g., see Xie (2001), Zhang et al. (2013), and McMahan et al. (2017). These authors demonstrate that by incorporating additional information, if available, one can obtain more efficient estimators and more precise inference. All of the aforementioned procedures are tailored to analyze data arising from specific group testing algorithms, with the work of McMahan et al. (2017) being the only methodology that offers a completely general framework.

The regression methodology we propose offers two main advantages over currently available methods. First, our framework can incorporate data arising from any group testing protocol, making it the most general frequentist-based procedure to date, with only McMahan et al. (2017) offering the same generality, but from a Bayesian perspective. Second, the proposed methodology makes use of a regularization technique of which the ridge (Hoerl and Kennard, 1970), lasso (Tibshirani, 1996), adaptive lasso (Zou, 2006), elastic net (Zou and Hastie, 2005), and adaptive elastic net (Zou and Zhang, 2009) are special cases. To our knowledge, regularized regression techniques have not yet been applied to group testing data.

In this paper we present an EM algorithm for computing a regularized logistic regression estimator for group testing data. The algorithm enables computation of the adaptive elastic net estimator, of which we establish theoretical properties in the group testing context. In particular, we show that it has an oracle property; i.e. as the sample size grows, the adaptive elastic net estimator will identify the true set of active covariates with probability tending to one, and it has the same asymptotic distribution as the estimator for which the true set of active covariates is known (by "active", we refer to covariates for which the regression coefficient is nonzero, and by "inactive", we refer to covariates for which the regression coefficient is equal to zero). Such properties are desirable to practitioners who employ group testing for diagnostic screening and disease surveillance.

2. Group testing and the log-likelihood

Let $Y_1, \dots, Y_N \in \{0, 1\}$ denote the true disease statuses of N individuals on which the covariates $X_1, \dots, X_N \in \mathbb{R}^p$ are observed and suppose that the conditional probability distribution of Y_i given X_i is

$$P_{\alpha,\beta}(Y_i \mid X_i) = \eta(\alpha + X_i^T \beta)^{Y_i}\{1 - \eta(\alpha + X_i^T \beta)\}^{1-Y_i}$$

for (α, β) equal to some $(\alpha_0, \beta_0) \in \mathbb{R} \times \mathbb{R}^p$, where η is a known link function (For ease of illustration we use the logit link $\eta(\cdot) = \exp(\cdot)/\{1 + \exp(\cdot)\}$, though our methodology can be easily generalized to other links, such as probit, as well). Moreover, suppose that the individual disease statuses $Y_1, \dots, Y_N$ are independent after conditioning on the covariates $X_1, \dots, X_N$. Our goal is to estimate (α_0, β_0) when instead of observing $Y_1, \dots, Y_N$ we observe data from a group testing procedure.

Generally, group testing data consists of collections of outcomes $\mathcal{A}_1, \dots, \mathcal{A}_J$ of assays taken on groups of individuals, where the groups are formed according to a partition $\mathcal{P}_1, \dots, \mathcal{P}_J$ of $\{1, \dots, N\}$. Each collection $\mathcal{A}_j$ contains outcomes of assays taken on pooled specimen of subsets of the individuals in $\mathcal{P}_j$. Assays may be taken on various subsets of the individuals in $\mathcal{P}_j$, including for subsets of size 1, and the collection of resulting outcomes comprises $\mathcal{A}_j$. Defining $\mathcal{X}_j = \{X_i, i \in \mathcal{P}_j\}$ and $\mathcal{Y}_j = \{Y_i, i \in \mathcal{P}_j\}$, we assume that the outcomes in $\mathcal{A}_j$ are liable to error with known sensitivities and specificities such that the conditional probability

$P(\mathcal{A}_j \mid \mathcal{Y}_j)$ of the assay outcomes from the individuals in group j given their true disease statuses is known and is free of the covariates $\mathcal{X}_j$ for all $j = 1, \ldots, J$; that is, the assay outcomes in $\mathcal{A}_j$ collected on the individuals in $\mathcal{P}_j$ are independent of $\mathcal{X}_j$ when conditioned on $\mathcal{Y}_j$.

In greater detail, for each $j = 1, \ldots, J$, define $\mathcal{P}_{j1}, \ldots, \mathcal{P}_{jL_j} \subseteq \mathcal{P}_j$ to be the subsets of individuals in $\mathcal{P}_j$ on which assays were conducted, where L_j is the total number of assays performed on subsets of the individuals in $\mathcal{P}_j$. Define the true disease statuses of the pools as $Z_{jl} = \max_{i \in \mathcal{P}_{jl}} Y_i$, for $l = 1, \ldots, L_j$, and let $\tilde{Z}_{j1}, \ldots, \tilde{Z}_{jL_j} \in \{0, 1\}$ be the assay results so that $\mathcal{A}_j = \{\tilde{Z}_{j1}, \ldots, \tilde{Z}_{jL_j}\}$. Since $Y_1, \ldots, Y_N$ are conditionally independent given $X_1, \ldots, X_N$ and $Z_{j1}, \ldots, Z_{jL_j}$ are conditionally independent given $\mathcal{Y}_j$, the assay results $\tilde{Z}_{j1}, \ldots, \tilde{Z}_{jL_j}$ are conditionally independent given the true pool statuses $Z_{j1}, \ldots, Z_{jL_j}$ and have Bernoulli distributions with success probabilities $\mathrm{Se}_{jl}^{Z_{jl}}(1 - \mathrm{Sp}_{jl})^{1-Z_{jl}}$, for $l = 1, \ldots, L_j$, where Se_{jl} and Sp_{jl} represent, respectively, the sensitivity and specificity of the test on the individuals in $\mathcal{P}_{jl}$. From here we have the expression

$$P(\mathcal{A}_j \mid \mathcal{Y}_j) = \prod_{l=1}^{L_j} \{\mathrm{Se}_{jl}^{Z_{jl}}(1 - \mathrm{Sp}_{jl})^{1-Z_{jl}}\}^{\tilde{Z}_{jl}} \{(1 - \mathrm{Se}_{jl})^{Z_{jl}} \mathrm{Sp}_{jl}^{1-Z_{jl}}\}^{1-\tilde{Z}_{jl}}. \quad (1)$$

From now on let $\mathcal{D}_N$ denote the observed group testing data, which is the set of independent collections of assay outcomes $\mathcal{A}_1, \ldots, \mathcal{A}_J$ and the covariate values $X_1, \ldots, X_N$.

Assuming a common density $f(\cdot)$ for $X_1, \ldots, X_N$, the log-likelihood based on $\mathcal{D}_N$ is

$$\ell(\alpha, \beta; \mathcal{D}_N) = \sum_{j=1}^{J} \left\{ \log P_{\alpha,\beta}(\mathcal{A}_j \mid \mathcal{X}_j) + \sum_{i \in \mathcal{P}_j} \log f(X_i) \right\},$$

where

$$P_{\alpha,\beta}(\mathcal{A}_j \mid \mathcal{X}_j) = \sum_{\mathrm{supp}\{\mathcal{Y}_j\}} P(\mathcal{A}_j \mid \mathcal{Y}_j) P_{\alpha,\beta}(\mathcal{Y}_j \mid \mathcal{X}_j), \quad (2)$$

where the summation is taken over the entire support $\mathrm{supp}\{\mathcal{Y}_j\}$ of the unobserved true disease statuses $\mathcal{Y}_j$, and where

$$P_{\alpha,\beta}(\mathcal{Y}_j \mid \mathcal{X}_j) = \prod_{i \in \mathcal{P}_j} \eta(\alpha + X_i^T \beta)^{Y_i} \{1 - \eta(\alpha + X_i^T \beta)\}^{1-Y_i}.$$

Figure 1 depicts three examples of group testing schemes.

[Figure 1 about here.]

Explicit expressions for $P(\mathcal{A}_j|\mathcal{Y}_j)$ under each of the group testing procedures depicted in Figure 1 are provided in Web Appendix A in the Supplementary Material. We restrict our remarks here to saying that master pool testing induces a likelihood in which the summation over $\text{supp}\{\mathcal{Y}_j\}$ in (2) admits a convenient simplification, but this is not the case for Dorfman or array testing. We note that for Dorfman testing $|\text{supp}\{\mathcal{Y}_j\}| = 2^{|\mathcal{P}_j|}$, so that if individuals are pooled into groups of size 10, $2^{10} = 1{,}024$ values must be summed to compute the contribution to the log-likelihood of one pool, which is not very burdensome. For array testing, however, $|\text{supp}\{\mathcal{Y}_j\}| = 2^{d_j \times d_j}$, where d_j is the array dimension, so that computing the contribution to the log-likelihood of a single 4 × 4 array would require summing over $2^{16} = 65{,}536$ values, and that of a single 5 × 5 array would require summing over $2^{25} = 33{,}554{,}432$ values.

In spite of the complicated form of the likelihoods induced by pooled testing procedures, the maximum likelihood estimator may be found with an EM algorithm in which the true unobserved disease statuses $Y_1, \ldots, Y_N$ are treated as missing data.

3. Penalized estimation under group testing

We consider penalized maximum likelihood estimators of the form

$$(\hat{\alpha}, \hat{\beta}) = \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\operatorname{argmin}} \; -\ell(\alpha, \beta; \mathcal{D}_N) + \lambda P_\theta^\omega(\beta), \tag{3}$$

where $\ell(\cdot, \cdot; \mathcal{D}_N)$ is the log-likelihood induced by the group testing procedure, and

$$P_\theta^\omega(\beta) = (1-\theta)\frac{1}{2}\sum_{j=1}^p \beta_j^2 + \theta \sum_{j=1}^p \omega_j |\beta_j|$$

with $\theta \in [0, 1]$ is a generalized version of the elastic net penalty (Zou, 2006; Zou and Zhang, 2009) with the weights $\omega_1, \ldots, \omega_p \in [0, \infty]$ applied to the ℓ_1 norm.

The form of the penalty is motivated by a sparsity assumption, i.e., the belief that not all the covariates are active in the true model. In particular, we assume that the set $S_0 = \{j : \beta_{0j} \neq 0\}$ of truly active covariates has cardinality $|S_0| < p$, so that the number of truly active covariates is less than the number p of covariates considered.

Writing $P_\theta^\omega(\beta)$ as $P_\theta(\beta)$ when ω is a p-length vector of ones, we define the maximum likelihood, elastic net, and adaptive elastic net estimators of (α_0, β_0) as

$$\begin{aligned}
(\hat{\alpha}^{\text{mle}}, \hat{\beta}^{\text{mle}}) &= \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{argmin}} \; -\ell(\alpha,\beta;\mathcal{D}_N) \\
(\hat{\alpha}^{\text{enet}}, \hat{\beta}^{\text{enet}}) &= \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{argmin}} \; -\ell(\alpha,\beta;\mathcal{D}_N) + \lambda P_\theta(\beta) \\
(\hat{\alpha}^{\text{aenet}}, \hat{\beta}^{\text{aenet}}) &= \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{argmin}} \; -\ell(\alpha,\beta;\mathcal{D}_N) + \lambda P_\theta^{\hat{\omega}}(\beta),
\end{aligned}$$

respectively, where we consider choosing $\hat{\omega}$ as

$$\hat{\omega}^{\text{enet}} = (|\hat{\beta}_1^{\text{enet}}|^{-\gamma}, \dots, |\hat{\beta}_p^{\text{enet}}|^{-\gamma}) \quad \text{or} \quad \hat{\omega}^{\text{mle}} = (|\hat{\beta}_1^{\text{mle}}|^{-\gamma}, \dots, |\hat{\beta}_p^{\text{mle}}|^{-\gamma}),$$

for some $\gamma > 0$. Each of these estimators are instances of the general estimator given in (3). Under the weights $\hat{\omega}^{\text{enet}}$ the elastic net estimator passes its sparsity to the adaptive elastic net estimator, as each covariate eliminated by the elastic net receives in the adaptive step a weight of $+\infty$. The weights $\hat{\omega}^{\text{mle}}$ encourage sparsity according to the magnitudes of the maximum likelihood coefficients. The elastic net and the adaptive elastic net estimators become the lasso and adaptive lasso estimators, respectively, when $\theta = 1$, and $\theta = 0$ corresponds to ridge regression. It is common to choose $\gamma = 1$ (Huang et al., 2008; van de Geer et al., 2011; Bühlmann and van de Geer, 2011), and we do so in our simulations and data analysis.

We remark that group testing data is intrinsically "large-N". Indeed, group testing procedures are used *because* the number of individuals N is large. The dimension p of the covariates $X_1, \dots, X_N$ is typically very small in comparison to N, so we are not concerned with a high-dimensional regime in which p exceeds (or grows with, in some sense) the number of individuals N. These penalized estimators, though they are in recent literature more and more associated with high-dimensional applications, are still of interest in low-dimensional settings for the sake of variable selection. But they offer another advantage: It may happen in binary regression that the unpenalized maximum likelihood estimator is undefined, due, for example, to complete separation or quasi-complete separation, under which the fitted coefficients diverge to ±∞ in order to achieve fitted probabilities of 0 or 1 (Albert and Anderson, 1984). When this is the case, penalization of the form in (3) can prevent the parameter estimates from diverging (Friedman et al., 2010).

4. EM algorithm for the penalized estimator

For any group testing procedure, we may express the log of the joint probability of the observed data $\mathcal{D}_N$ and the set of unobserved true individual statuses $\mathcal{Y} = \{Y_1, \ldots, Y_N\}$ as

$$\tilde{\ell}(\alpha, \beta; \mathcal{D}_N, \mathcal{Y}) = \sum_{i=1}^{N} \log P_{\alpha,\beta}(Y_i \mid X_i) + C(\mathcal{D}_N, \mathcal{Y}),$$

where $C(\mathcal{D}_N, \mathcal{Y})$ depends on the testing procedure but not on (α, β), and where

$$\log P_{\alpha,\beta}(Y_i \mid X_i) = Y_i \log \eta(\alpha + X_i^T \beta) + (1 - Y_i) \log\{1 - \eta(\alpha + X_i^T \beta)\}.$$

Assuming that we may compute the conditional expections $\mathbb{E}_{\alpha,\beta}(Y_i \mid \mathcal{D}_N)$ for $i = 1, \ldots, N$, where $\mathbb{E}_{\alpha,\beta}$ denotes expectation under $(\alpha_0, \beta_0) = (\alpha, \beta)$, the maximum likelihood estimator $(\hat{\alpha}^{\text{mle}}, \hat{\beta}^{\text{mle}})$ can be found via the EM algorithm through the updates

$$(\alpha, \beta)^{(k+1)} := \underset{(\alpha,\beta) \in \mathbb{R} \times \mathbb{R}^p}{\operatorname{argmax}} \; \mathbb{E}_{(\alpha,\beta)^{(k)}} \left\{ \sum_{i=1}^{N} \log P_{\alpha,\beta}(Y_i \mid X_i) \mid \mathcal{D}_N \right\}, \tag{4}$$

starting from some initial value $(\alpha, \beta)^{(0)} \in \mathbb{R} \times \mathbb{R}^p$, where computing the conditional expectation in (4) involves only computing the $\mathbb{E}_{(\alpha,\beta)^{(k)}}(Y_i | \mathcal{D}_N)$ for $i = 1, \ldots, N$.

Moreover, we can compute the penalized maximum likelihood estimator from (3) by applying the elastic net penalty to each update in (4). This gives the EM algorithm updates

$$(\alpha, \beta)^{(k+1)} := \underset{(\alpha,\beta) \in \mathbb{R} \times \mathbb{R}^p}{\operatorname{argmax}} \; \mathbb{E}_{(\alpha,\beta)^{(k)}} \left\{ \sum_{i=1}^{N} \log P_{\alpha,\beta}(Y_i \mid X_i) \mid \mathcal{D}_N \right\} - \lambda P_\theta^\omega(\beta),$$

starting from some initial value $(\alpha, \beta)^{(0)} \in \mathbb{R} \times \mathbb{R}^p$.

Each update of the penalized EM-algorithm may be computed via the coordinate descent algorithm of Friedman et al. (2010); in each update, we simply compute an elastic-net penalized logistic regression estimator where the responses are the conditional expectations of $Y_1, \ldots, Y_N$ at the current parameter values. This is summarized in Algorithm 1, in which we describe the EM algorithm for maximizing (3) for any choice of θ, λ, and ω. Algorithm S.1 in Web Appendix B of the Supplementary Material gives the complete details of the coordinate descent algorithm. Web Appendix C of the Supplementary Material discusses computing the elastic net and adaptive elastic net estimators over many values of the tuning parameter $\lambda \in [0, \infty]$ with "warm starts" to speed up computation.

```
Data: pooled testing data $\mathcal{D}_N$, initial value $(\alpha, \beta)^{(0)} \in \mathbb{R} \times \mathbb{R}^p$, stopping criterion $\delta$,
      tuning parameters $\lambda$, $\theta$, and $\omega_1, \ldots, \omega_p$.
Result: $\underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{argmin}} \; -\ell(\alpha, \beta; \mathcal{D}_N) + \lambda P_\theta^\omega(\beta)$
$\Delta^\dagger \longleftarrow \delta + 1$
$(\alpha, \beta)^\dagger \longleftarrow (\alpha, \beta)^{(0)}$
while $\Delta^\dagger > \delta$ do
    $(\alpha, \beta)^\ddagger \longleftarrow (\alpha, \beta)^\dagger$
    $Y_i^* \longleftarrow \mathbb{E}_{(\alpha,\beta)^\ddagger}(Y_i \mid \mathcal{D}_N)$, $i = 1, \ldots, N$
    $(\alpha, \beta)^\dagger \longleftarrow \underset{(\alpha,\beta)\in\mathbb{R}\times\mathbb{R}^p}{\text{argmax}} \sum_{i=1}^N Y_i^* \log \eta(\alpha + X_i^T\beta) + (1 - Y_i^*) \log\{1 - \eta(\alpha + X_i^T\beta)\} - \lambda P_\theta^\omega(\beta)$
    $\Delta^\dagger \longleftarrow \max\{|\alpha^\dagger - \alpha^\ddagger|, |\beta_1^\dagger - \beta_1^\ddagger|, , \ldots, |\beta_p^\dagger - \beta_p^\ddagger|\}$
$(\hat{\alpha}, \hat{\beta}) \longleftarrow (\alpha, \beta)^\dagger$
```

Algorithm 1: EM-algorithm to compute the penalized estimator of (3).

We note that the group testing procedure itself enters the EM-algorithm only in the computation of the conditional expectations $\mathbb{E}_{\alpha,\beta}(Y_i \mid \mathcal{D}_N)$, for $i = 1, \ldots, N$.

For any group testing procedure, we have

$$\mathbb{E}_{\alpha,\beta}(Y_i \mid \mathcal{D}_N) = \frac{P_{\alpha,\beta}(\mathcal{A}_j \mid Y_i, \mathcal{X}_j) P_{\alpha,\beta}(Y_i \mid X_i)\big|_{Y_i=1}}{\sum_{y_i=0}^1 P_{\alpha,\beta}(\mathcal{A}_j \mid Y_i, \mathcal{X}_j) P_{\alpha,\beta}(Y_i \mid X_i)\big|_{Y_i=y_i}},$$

for $i \in \mathcal{P}_j$, and in general

$$P_{\alpha,\beta}(\mathcal{A}_j \mid Y_i, \mathcal{X}_j) = \sum_{\text{supp}\{\mathcal{Y}_j^{(-i)}\}} P(\mathcal{A}_j \mid Y_i, \mathcal{Y}_j^{(-i)}) P_{\alpha,\beta}(\mathcal{Y}_j^{(-i)} \mid \mathcal{X}_j^{(-i)}),$$

where $\mathcal{Y}_j^{(-i)} = \{Y_{i'}, i \neq i' \in \mathcal{P}_j\}$ and $\mathcal{X}_j^{(-i)} = \{X_{i'}, i \neq i' \in \mathcal{P}_j\}$.

We may compute $\mathbb{E}_{\alpha,\beta}(Y_i \mid \mathcal{D}_N)$ exactly under master pool or Dorfman testing at no great computational cost. Under array testing, however, we recommend an MCMC approximation to $\mathbb{E}_{\alpha,\beta}(Y_i|\mathcal{D}_N)$ which we describe in Web Appendix D of the Supplementary Material.

5. Technical results

In this section we give results concerning the behavior of the adaptive elastic net estimator under mild assumptions. We have placed our assumptions, (A.1), (A.2), and (A.3), as well as the complete proofs of the results that follow in Web Appendix E of the Supplementary Material, and we discuss them here only briefly. Assumption (A.1) ensures that as N increases, so does the number of groups of individuals and thus the number of independent contributions to the likelihood. This allows us to frame our asymptotics in terms of the number of individuals N. Assumption (A.2) guarantees that the asymptotic covariance matrix of our estimator is positive definite. It generally holds if the assay has a nontrivial diagnostic ability; i.e. the sum of the assay sensitivity and specificity is not equal to 1. Assumption (A.3) guarantees bounded third partial derivatives of the log-likelihood function. Both Assumptions (A.2) and (A.3) are standard in deriving the asymptotic normality of a maximum likelihood estimator; e.g., see Lehmann and Casella (2006).

We begin with the following preliminary result concerning the behavior of the non-adaptive elastic net and the maximum likelihood estimator. The result tells us that both of these estimators are suitable for defining the weights for the adaptive estimator.

Lemma 1: *Under Assumptions (A.1), (A.2), and (A.3) and for any* $\theta \in [0,1]$, *if* $\lambda N^{-1/2} \to 0$ *as* $N \to \infty$, *then there exists a local minimizer of* $-\ell(\alpha, \beta; \mathcal{D}_N) + \lambda P_\theta(\beta)$, *denoted by*

$(\tilde{\alpha}, \tilde{\beta}^T)^T$, *such that*

$$|\tilde{\alpha} - \alpha_0| = O_p(N^{-1/2}) \text{ and } \|\tilde{\beta} - \beta_0\| = O_p(N^{-1/2}). \tag{5}$$

Remark 1: When $\lambda = 0$, $\tilde{\alpha} = \hat{\alpha}^{\text{mle}}$ and $\tilde{\beta} = \hat{\beta}^{\text{mle}}$; when $\lambda > 0$, $\tilde{\alpha} = \hat{\alpha}^{\text{enet}}$ and $\tilde{\beta} = \hat{\beta}^{\text{enet}}$.

To state our main result concerning the adaptive elastic net estimator, the proof of which makes use of Lemma 1, we require the following notation: For any set $S \subset \{1, \ldots, p\}$, denote by β_S the vector formed by keeping the entries in β with indices in S. Moreover, let $\mathcal{I}(\alpha_0, \beta_0)$ be the information matrix corresponding the log-likelihood $\ell(\alpha, \beta; \mathcal{D}_N)$ evaluated at the true parameter values (α_0, β_0), and let $\mathcal{I}_{S_0,S_0}(\alpha_0, \beta_0)$ be the submatrix of $\mathcal{I}(\alpha_0, \beta_0)$ formed by keeping the rows and columns with indices in S_0.

Theorem 1: *Under Assumptions (A.1), (A.2), and (A.3) and for any* $\theta \in (0, 1]$, *if* $\lambda N^{-1/2} \to 0$ *and* $\lambda N^{(\gamma-1)/2} \to \infty$ *as* $N \to \infty$, *then*

(1) $P(\hat{\beta}^{\text{aenet}}_{S_0^c} = 0) \to 1$

(2) $$\sqrt{N}\left(\begin{bmatrix} \hat{\alpha}^{\text{aenet}} \\ \hat{\beta}^{\text{aenet}}_{S_0} \end{bmatrix} - \begin{bmatrix} \alpha_0 \\ \beta_{0S_0} \end{bmatrix}\right) \to N\{0, \mathcal{I}_{S_0,S_0}(\alpha_0, \beta_0)^{-1}\} \qquad \textit{in distribution.}$$

Remark 2: Theorem 1 holds for either choice of weights $\hat{\omega}^{\text{mle}}$ and $\hat{\omega}^{\text{enet}}$ in the construction of the adaptive elastic net estimator. Note that Lemma 1 shows $\|\hat{\beta}^{\text{enet}} - \beta_0\| = O_p(N^{-1/2})$ or $\|\hat{\beta}^{\text{mle}} - \beta_0\| = O_p(N^{-1/2})$, which is the key in Theorem 1. For inactive covariates, the weights approach infinity, whereas the weights for active covariates converge to finite constants. Thus, with a good choice of tuning parameters (θ, λ), where $\theta > 0$, the estimator $\hat{\beta}^{\text{aenet}}$ would (asymptotically) identify the set of inactive covariates and estimate the nonzero coefficients without bias. On the other hand, if $\theta = 0$, then selection consistency; i.e., $P(\hat{\beta}^{\text{aenet}}_{S_0^c} = 0) \to 1$, cannot be achieved because the lasso-type penalty vanishes. However, if we keep $\lambda N^{-1/2} \to$

0 as $N \to \infty$, we still have a root-N consistent estimator; i.e., $\sqrt{N}\{(\hat{\alpha}^{\text{aenet}}, \hat{\beta}^{\text{aenet}\,T})^T - (\alpha_0, \beta_0^T)^T\} \to N\{0, \mathcal{I}(\alpha_0, \beta_0)^{-1}\}$ in distribution.

Remark 3: Even though the adaptive elastic net estimators of the nonzero coefficients are asymptotically Normal, there is empirical work which suggests that the convergence is quite slow, and that Wald-type confidence intervals for the coefficients in β_{0S_0} will have poor coverage. For example, it was demonstrated in Das et al. (2017) that the values of λ which ensure good variable selection properties (larger values of λ) tend to result in subnominal coverage of Wald-type intervals; smaller values of λ, under which variable selection performance is poor, result in closer-to-nominal coverage of Wald-type intervals. We emphasize that it is the variable selection result, statement (1) of Theorem 1, which is of primary interest in regularized regression modeling, while the asymptotic Normality result, statement (2) of Theorem 1, comes as a theoretical byproduct and is seldom used by practitioners to conduct Wald-type inference.

6. Selection of tuning parameters

For testing procedures under which the log-likelihood can be computed, many standard model comparison criteria are available for choosing λ and θ, such as the BIC or the ERIC criterion from Hui et al. (2015). It is not guaranteed that these criteria will select a sequence of λ values which satisfies the conditions of Theorem 1 as $N \to \infty$, but they offer, for a fixed sample size N, a reasonable way to compare model fits at different values of λ and θ, allowing these to be chosen from the data.

For the estimator $(\hat{\alpha}, \hat{\beta})$ minimizing (3), we define the BIC and ERIC criteria as

$$\text{BIC}(\hat{\alpha}, \hat{\beta}) = -2\ell(\hat{\alpha}, \hat{\beta}; \mathcal{D}_N) + \widehat{\text{df}}(\hat{\alpha}, \hat{\beta}) \log(N)$$

$$\text{ERIC}(\hat{\alpha}, \hat{\beta}) = -2\ell(\hat{\alpha}, \hat{\beta}; \mathcal{D}_N) + \widehat{\text{df}}(\hat{\alpha}, \hat{\beta}) \log(N/\lambda),$$

where the degrees of freedom $\mathrm{df}(\hat{\alpha}, \hat{\beta})$ of the estimator $(\hat{\alpha}, \hat{\beta})$ is estimated by

$$\widehat{\mathrm{df}}(\hat{\alpha}, \hat{\beta}) = 1 + \mathrm{tr}\left[\Xi_{\hat{S}}\left\{\Xi_{\hat{S}}^T \hat{W} \Xi_{\hat{S}} + \lambda(1-\theta) I_{|\hat{S}|}\right\}^{-1} \Xi_{\hat{S}}^T \hat{W}\right],$$

where $\hat{S} = \{j \in \{1, \ldots, p\} : \hat{\beta}_j \neq 0\}$, $\Xi_{\hat{S}}$ is the matrix containing the columns $\hat{S}$ of the $N \times p$ design matrix $\Xi = (X_1^T, \ldots, X_N^T)^T$, and $\hat{W}$ is a diagonal matrix with diagonals entries $\hat{w}_{ii} = \eta(\hat{\alpha} + X_i^T \hat{\beta})\{1 - \eta(\hat{\alpha} + X_i^T \hat{\beta})\}$, $i = 1, \ldots, N$ (Tibshirani and Taylor, 2012).

We remark that under master pool and Dorfman testing the log-likelihood $\ell(\hat{\alpha}, \hat{\beta}; \mathcal{D}_N)$ can be computed rather easily, provided that the pool size under Dorfman testing is not too large. Under array testing, however, since $\mathrm{supp}\{\mathcal{Y}_j\}$ is typically too large to allow exact computation of the log-likelihood, we choose to approximate the log-likelihood contribution $P_{\alpha,\beta}(\mathcal{A}_j \mid \mathcal{X}_j)$ of the jth array by the Monte Carlo approximation $B^{-1}\sum_{b=1}^B P(\mathcal{A}_j \mid \{Y_i^{(b)}, i \in \mathcal{P}_j\})$, where $Y_i^{(b)} \sim \mathrm{Bernoulli}\{\eta(\alpha + X_i^T\beta)\}$, independent, for $i \in \mathcal{P}_j$, $b = 1, \ldots, B$, with B large, by way of importance sampling.

An alternative to using the BIC or ERIC criteria is to choose the tuning parameters via a likelihood-based crossvalidation procedure in which pools of individuals are removed in order to obtain training and testing folds. We describe a crossvalidation procedure for group testing in detail in Web Appendix F of the Supplementary Material.

For the adaptive elastic net estimator under the weights $\hat{\omega}^{\mathrm{mle}}$, the tuning parameter needs only to be chosen once, in the adaptive step. However, under the weights $\hat{\omega}^{\mathrm{enet}}$, a pair of tuning parameters (λ, θ) must also be selected for the initial estimator. Each of the methods discussed above, BIC, ERIC, and crossvalidation, can be used in selecting the tuning parameters for the initial estimator and then used a second time for selecting the tuning parameters for the adaptive estimator.

7. Analysis of Iowa Chlamydia Data

In this section, the proposed methodology is used to analyze chlamydia data collected by the State Hygienic Laboratory (SHL) in Iowa City during the 2014 calendar year. The current screening protocols implemented by the SHL require that all male specimen and female urine specimen be tested individually. In contrast, female swab specimen are tested in pools (usually of size 4), with positive pools being resolved through individual level testing; for further discussion of the screening protocol implemented by the SHL see McMahan et al. (2017) and the references therein. Thus, this analysis focuses solely on the test results that were collected on the N =13,862 female subjects screened during the 2014 calendar year.

The available data consists of test results taken on 2,273 swab master pools of size 4, 12 swab master pools of size 3, 1 swab master pool of size 2, 416 individual swab specimen, and 4,316 individual urine specimen, as well as the test results required to resolve positive master pools. There are 10 covariates: age in years, race indicators (Caucasian, African American, and other), sexual practice indicators (a new sexual partner was reported in the last 90 days, multiple partners were reported in the last 90 days), a risk indicator (the individual had contact with a partner having any sexually transmitted disease reported in the previous year), clinical symptom indicators (the individual presented with common symptoms of infection, cervicitis, pelvic inflammatory disease), and a specimen type indicator.

[Figure 2 about here.]

Figure 2 displays the solution paths for the elastic net and adaptive elastic net under the weights $\hat{\omega}^{\text{mle}}$ across 30 values of the tuning parameter λ for the θ values $0, 1/8, 1/4, 1/2$ and 1, where $\theta = 0$ corresponds to ridge regression and $\theta = 1$ corresponds to the lasso. The open symbols in Figure 2 trace the solution path of the elastic net estimator and the filled symbols that of the adaptive elastic net estimator. Note that when $\theta = 0$, the adaptive part of the penalty in (3) is removed, as it affects only the ℓ_1 norm, so there is no adaptive estimator.

Also indicated in Figure 2 are the tuning parameter choices made from the 30 × 5 grid of candidate (λ, θ) pairs for the elastic net and adaptive elastic net under the weights $\hat{\omega}^{\mathrm{mle}}$ by 5-fold crossvalidation and by the BIC and ERIC criteria. Table 1 displays the values of the adaptive elastic net estimator under the weights $\hat{\omega}^{\mathrm{mle}}$ as well as the values of the non-adaptive elastic net estimator and the unpenalized maximum likelihood estimator. The tuning parameter selections for each estimator are also shown; when the weights $\hat{\omega}^{\mathrm{enet}}$ are used, the tuning parameters are selected twice, once for the initial estimator and once for the adaptive estimator. When the weights $\hat{\omega}^{\mathrm{mle}}$ are used, fewer variables are eliminated. Table 1 also shows for which coefficients the maximum likelihood estimator had a p-value less than 0.01 (indicated by asterisks) for the equal-to-zero null hypothesis. The p-values of the maximum likelihood coefficients are computed with respect to a Normal distribution with mean 0 and covariance matrix given by the inverse of the observed information as computed via Louis' method (see Web Appendix G of the Supplementary Material). If variables were identified as active on this basis, the selections would agree with those of the adaptive elastic net under the crossvalidation, BIC, and ERIC tuning parameter selection methods when the elastic net weights $\hat{\omega}^{\mathrm{enet}}$ are used. In Web Appendix H of the Supplementary Material we provide a plot of the solution paths computed on each of the 5 crossvalidation training sets.

Reassuringly, the results from this analysis are in agreement with previous epidemiological knowledge of chlamydia infection in females (e.g., see Navarro et al., 2003). In particular, all of the considered regression methodologies identified an increase risk of chlamydia infection being attributable to having a new sexual partner within the last 90 days, having multiple partners during the last 90 days, having had contact with a partner having any sexually transmitted disease reported in the previous year, and having common symptoms of infection. Interestingly, evidence of cervicitis and pelvic inflammatory disease do not seem to be important in the presence of the other variables, likely because these risk factors are accounted for through the symptom indicator. When compared to other ethnic backgrounds, Caucasian (African American) females appear to be at lower (higher) risk of chlamydia infection. Lastly, risk seems to diminish with age, while the specimen type does not appear to be related. In conclusion, in this analysis the proposed regression methodologies generally provided for the same conclusions that would be obtained under the maximum likelihood approach. This is not always the case as is demonstrated through numerical simulation in the following section.

[Table 1 about here.]

8. Simulation studies

We study via simulation the variable selection, prediction, and estimation performance of the adaptive elastic net estimator. We compare its predictive and estimation performance to those of the unpenalized maximum likelihood estimator and of the oracle estimator, which is the unpenalized maximum likelihood estimator computed only considering the set of truly active covariates, that is, only the covariates in $S_0 = \{j : \beta_{0j} \neq 0\}$.

To simulate group testing data we generate individual covariate observations $X_1, \dots, X_N$ as independent realizations of a random variable X from a multivariate normal distribution with some covariance matrix Σ. Then we generate true disease statuses $Y_1, \dots, Y_N$ as independent Bernoulli random variables with success probabilities $\eta(\alpha_0 + X_i^T \beta_0)$, $i = 1, \dots, N$. We generate assay results $\mathcal{A}_1, \dots, \mathcal{A}_J$ from $Y_1, \dots, Y_N$ according to a group testing procedure, and we compute our estimators based on the observed data $X_1, \dots, X_N$ and $\mathcal{A}_1, \dots, \mathcal{A}_J$. In all simulations, pools were formed at random, i.e. without regard to covariate values.

We consider the following models for generating the covariate values and disease statuses: *Model 1*: $\alpha_0 = -4$, $\beta_0 = (2/3, 1/3, 1, 0 \cdot 1_7^T)^T$, with $\Sigma = 0.5 \cdot 1_{10} 1_{10}^T + 0.5 \cdot I_{10}$, where I_d is the $d \times d$ identity matrix and 1_d is a $d \times 1$ vector of ones, under which the disease prevalence is

$\mathbb{E}\eta(\alpha_0 + X^T\beta_0) \approx 0.051$. This model has 3 active covariates of 10 and the correlation between all pairs of covariates is 0.5.

Model 2: $\alpha_0 = -4$, $\beta_0 = \{(-3/4, -1/2, -1/4, 1/4, 1/2, 3/4)^T, 0{\cdot}1_{12}^T\}^T$, with $\Sigma = \{2^{-|l-l'|}\}_{1\leqslant l,l'\leqslant 18}$ under which the disease prevalence is $\mathbb{E}\eta(\alpha_0 + X^T\beta_0) \approx 0.082$. There are 6 active covariates of 18 and all pairs of covariates have correlations of various strengths.

Model 3: $\alpha_0 = -3$, $\beta_0 = \{1_8^T \otimes (1/2, 0, 0)^T\}^T$, with $\Sigma = I_8 \otimes \{(9/10)^{|l-l'|}\}_{1\leqslant l,l'\leqslant 3}$, under which the disease prevalence is $\mathbb{E}\eta(\alpha_0 + X^T\beta_0) \approx 0.092$. There are 6 active covariates of 24, and active covariates are independent from each other but each active covariate is highly correlated with two inactive covariates.

We consider estimating (α_0, β_0) from each of the above models under master pool, Dorfman, and array testing, where individuals are grouped into pools of size 5 under master pool and Dorfman testing and into 5 × 5 arrays under array testing. We also consider estimating (α_0, β_0) when individual testing instead of group testing is used. We consider sample sizes of $N = 1{,}000$ and $N = 5{,}000$ individuals. To assays on pools of more than one individual under any group testing procedure, we assign the sensitivity 0.92 and specificity 0.96 and to assays on single individuals we assign the sensitivity 0.95 and the specificity 0.98. We consider both choices of the weights $\hat{\omega}^{\text{enet}}$ and $\hat{\omega}^{\text{mle}}$ which may be used to construct the adaptive elastic net estimator. When the weights $\hat{\omega}^{\text{enet}}$ are used, λ and θ are chosen separately for the initial estimator and for the adaptive estimator. We consider selecting θ and λ by crossvalidation, BIC, and ERIC as described in Section 6, considering a 30 × 3 grid of 30 candidate λ values for each θ, with candidate θ values of 1/4, 1/2, and 1.

Table 2 gives Monte Carlo estimates under $N = 10{,}000$ of the quantity

$$\mathbb{E}[\mathbb{E}\{\eta(\alpha_0 + X^T\beta_0) - \eta(\hat{\alpha} + X^T\hat{\beta})|\hat{\alpha}, \hat{\beta}\}^2]^{1/2} \times 100, \tag{6}$$

which represents the expected error when predicting the probability of disease for a randomly selected individual, scaled by 100 to render it in percentage points. The inner expectation is approximated via 10,000 Monte Carlo draws of X and the outer expectation is estimated from 500 simulated data sets (standard errors are given in parentheses). For each model and each testing procedure we present in bold face the highest and the two lowest expected prediction errors across all estimators (apart from the oracle, which always achieves the lowest expected prediction error); the maximum likelihood estimator is the worst performer under every model and assay scheme. Results are similar when we consider the mean estimation error $\mathbb{E}\|(\hat{\alpha}, \hat{\beta}^T)^T - (\alpha_0, \beta_0^T)^T\|$. These results as well as all $N = 1{,}000$ results are provided in Tables S.1–S.4 and Figures S.2–S.13 in Web Appendix I of the Supplementary Material.

We provide also in the "procedure" column of Table 2 the average number of assays performed on the sets of $N = 5{,}000$ individuals under each group testing scheme. This is fixed at 1,000 under master pool testing with master pools of size 5 and fixed at 5,000 under individual testing. Under Dorfman and array testing, however, the total number of assays is random. The expected prediction errors achieved under Dorfman and array testing are very close to, and in some cases better than those achieved under individual testing, even though they require many fewer assays; this highlights the potential of group testing to reduce costs without compromising estimation performance.

[Table 2 about here.]

To give a sense of the variable selection performance of the adaptive elastic net estimator we depict in Figure 3 the frequencies with which each covariate was selected under Model 1 when Dorfman testing was being utilized and the tuning parameters were being selected via crossvalidation, BIC, and ERIC. The bottom rows of the figure display the proportion of times the selected set of covariates contained and was equal to the true set of active covariates. For the larger sample size $N = 5{,}000$, each relevant covariate is selected with greater frequency and the true set of active covariates is more often selected, heuristically demonstrating the selection consistency of the adaptive elastic net estimator under these data-based choices of the tuning parameters. Further, we see that the estimator is effective under both weight choices $\hat{\omega}^{\text{enet}}$ and $\hat{\omega}^{\text{mle}}$.

[Figure 3 about here.]

In addition to these studies, we also conducted a robustness study to assess the adaptive elastic net estimator when it is fit using incorrect values of sensitivity and specificity. Our results, summarized in Table S.5 in Web Appendix J of the Supplementary Material, show that the adaptive elastic net is still a reliable estimator under moderate misspecification of these quantities. We also conducted a simulation study comparing the performance of the non-adaptive elastic net and the adaptive elastic net for variable selection as the sample size is increased; Figure S.14 in Web Appendix K of the Supplementary Material shows that the adaptive estimator achieved much better results, as the theory suggests.

Acknowledgement

This research was partially supported by National Institutes of Health grant AI121351.

9. Supplementary Materials

Web Appendices, Tables, and Figures referenced in Sections 2-8 as well as a zip file with software and examples are available with this paper at the *Biometrics* website on Wiley Online Library.

References

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models. *Biometrika* **71,** 1–10.

Bühlmann, P. and van de Geer, S. (2011). *Statistics for High-Dimensional Data. Methods, Theory and Applications.* Springer, Heidelberg.

Chen, P., Tebbs, J. M., and Bilder, C. R. (2009). Group testing regression models with fixed and random effects. *Biometrics* **65,** 1270–1278.

Das, D., Gregory, K., and Lahiri, S. (2017). Perturbation bootstrap in adaptive lasso. *arXiv preprint arXiv:1703.03165* .

Delaigle, A. and Hall, P. (2015). Nonparametric methods for group testing data, taking dilution into account. *Biometrika* **102,** 871–887.

Delaigle, A., Hall, P., and Wishart, J. (2014). New approaches to non-and semi-parametric regression for univariate and multivariate group testing data. *Biometrika* **101,** 567–585.

Delaigle, A. and Meister, A. (2011). Nonparametric regression analysis for group testing data. *Journal of the American Statistical Association* **106,** 640–650.

Dorfman, R. (1943). The detection of defective members of large populations. *The Annals of Mathematical Statistics* **14,** 436–440.

Farrington, C. (1992). Estimating prevalence by group testing using generalized linear models. *Statistics in medicine* **11,** 1591–1597.

Friedman, J., Hastie, T., and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. *Journal of statistical software* **33,** 1–22.

Gastwirth, J. L. and Johnson, W. O. (1994). Screening with cost-effective quality control: potential applications to hiv and drug testing. *Journal of the American Statistical Association* **89,** 972–981.

Heffernan, A. L., Aylward, L. L., Leisa-maree, L., Sly, P. D., Macleod, M., and Mueller, J. F. (2014). Pooled biological specimens for human biomonitoring of environmental chemicals: opportunities and limitations. *Journal of Exposure Science and Environmental Epidemiology* **24,** 225–232.

Hoerl, A. E. and Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics* **12,** 55–67.

Huang, J., Ma, S., and Zhang, C.-H. (2008). Adaptive lasso for sparse high-dimensional regression models. *Statistica Sinica* **18,** 1603–1618.

Huang, X. (2009). An improved test of latent-variable model misspecification in structural measurement error models for group testing data. *Statistics in medicine* **28,** 3316–3327.

Hui, F. K. C., Warton, D. I., and Foster, S. D. (2015). Tuning parameter selection for the adaptive lasso using eric. *Journal of the American Statistical Association* **110,** 262–269.

Kim, H.-Y., Hudgens, M. G., Dreyfuss, J. M., Westreich, D. J., and Pilcher, C. D. (2007). Comparison of group testing algorithms for case identification in the presence of test error. *Biometrics* **63,** 1152–1163.

Krajden, M., Cook, D., Mak, A., Chu, K., Chahil, N., Steinberg, M., Rekart, M., and Gilbert, M. (2014). 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. *Journal of Clinical Virology* **61,** 132–137.

Lehmann, E. L. and Casella, G. (2006). *Theory of point estimation.* Springer Science & Business Media.

Lewis, J. L., Lockary, V. M., and Kobic, S. (2012). Cost savings and increased efficiency using a stratified specimen pooling strategy for chlamydia trachomatis and neisseria gonorrhoeae. *Sexually transmitted diseases* **39,** 46–48.

Liu, A., Liu, C., Zhang, Z., and Albert, P. S. (2011). Optimality of group testing in the presence of misclassification. *Biometrika* **99,** 245–251.

McMahan, C. S., Tebbs, J. M., and Bilder, C. R. (2012). Regression models for group testing data with pool dilution effects. *Biostatistics* **14,** 284–298.

McMahan, C. S., Tebbs, J. M., Hanson, T. E., and Bilder, C. R. (2017). Bayesian regression for group testing data. *Biometrics* **73,** 1443–1452.

Navarro, C., Jolly, A., Nair, R., and Chen, Y. (2003). Risk factors for genital chlamydial infection. *Journal of Sexual & Reproductive Medicine* **3,** 23–34.

Thompson, K. H. (1962). Estimation of the proportion of vectors in a natural population of insects. *Biometrics* **18,** 568–578.

Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological)* pages 267–288.

Tibshirani, R. J. and Taylor, J. (2012). Degrees of freedom in lasso problems. *The Annals of Statistics* **40,** 1198–1232.

van de Geer, S., Bhlmann, P., and Zhou, S. (2011). The adaptive and the thresholded lasso for potentially misspecified models (and a lower bound for the lasso). *Electron. J. Statist.* **5,** 688–749.

Vansteelandt, S., Goetghebeur, E., and Verstraeten, T. (2000). Regression models for disease prevalence with diagnostic tests on pools of serum samples. *Biometrics* **56,** 1126–1133.

Wang, D., McMahan, C., Gallagher, C., and Kulasekera, K. (2014). Semiparametric group testing regression models. *Biometrika* **101,** 587–598.

Xie, M. (2001). Regression analysis of group testing samples. *Statistics in medicine* **20,** 1957–1969.

Zhang, B., Bilder, C. R., and Tebbs, J. M. (2013). Group testing regression model estimation when case identification is a goal. *Biometrical Journal* **55,** 173–189.

Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American statistical association* **101,** 1418–1429.

Zou, H. and Hastie, T. (2005). Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* **67,** 301–320.

Zou, H. and Zhang, H. H. (2009). On the adaptive elastic-net with a diverging number of parameters. *Annals of statistics* **37,** 1733–1751.

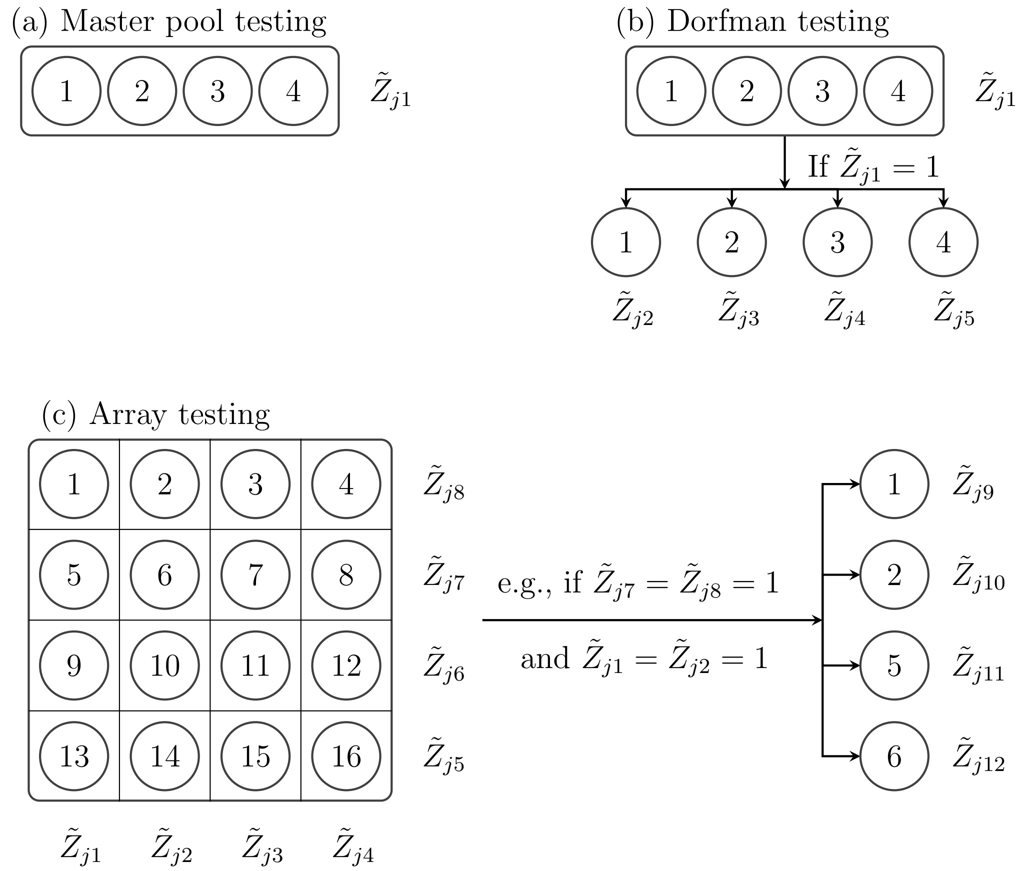


Figure 1. Illustration of three group testing procedures: (a) Master pool testing, the pooled specimen (formed by combining the specimen collected from the individuals identified by $\mathcal{P}_j = \{1, 2, 3, 4\}$) is assayed yielding $\tilde{Z}_{j1}$ so that $\mathcal{A}_j = \{\tilde{Z}_{j1}\}$. (b) Dorfman testing, first stage proceeds identically to master pool testing. If the master pool tests negative, then $\tilde{Z}_{j1} = 0$ and $\mathcal{A}_j = \{\tilde{Z}_{j1} = 0\}$; if it tests positive, each of the individuals is retested individually and $\mathcal{A}_j = \{\tilde{Z}_{j1} = 1, \tilde{Z}_{j2}, \tilde{Z}_{j3}, \tilde{Z}_{j4}, \tilde{Z}_{j5}\}$, where $\tilde{Z}_{j2}, \tilde{Z}_{j3}, \tilde{Z}_{j4}, \tilde{Z}_{j5}$ are the individual level testing outcomes. (c) Array testing, the individuals in $\mathcal{P}_j = \{1, 2, \ldots, 16\}$ are assigned to a 4×4 array. Row and column pools are formed and assayed yielding four column outcomes $\tilde{Z}_{j1}, \ldots, \tilde{Z}_{j4}$ and four row outcomes $\tilde{Z}_{j5}, \ldots, \tilde{Z}_{j8}$. If an individual's row and column pools are both positive then he/she is retested individually; if there are no such individuals in an array, all individuals in the array belonging to a single positive pool are retested; see Kim et al. (2007). In this example, the first two rows and the first two columns test positive, so that individuals 1, 2, 5, and 6 are retested, producing the individual level testing outcomes $\tilde{Z}_{j9}, \tilde{Z}_{j10}, \tilde{Z}_{j11}, \tilde{Z}_{j12}$. In this case $\mathcal{A}_j = \{\tilde{Z}_{j8} = \tilde{Z}_{j7} = \tilde{Z}_{j1} = \tilde{Z}_{j2} = 1, \tilde{Z}_{j3} = \tilde{Z}_{j4} = \tilde{Z}_{j5} = \tilde{Z}_{j6} = 0, \tilde{Z}_{j9}, \tilde{Z}_{j10}, \tilde{Z}_{j11}, \tilde{Z}_{j12}\}$.

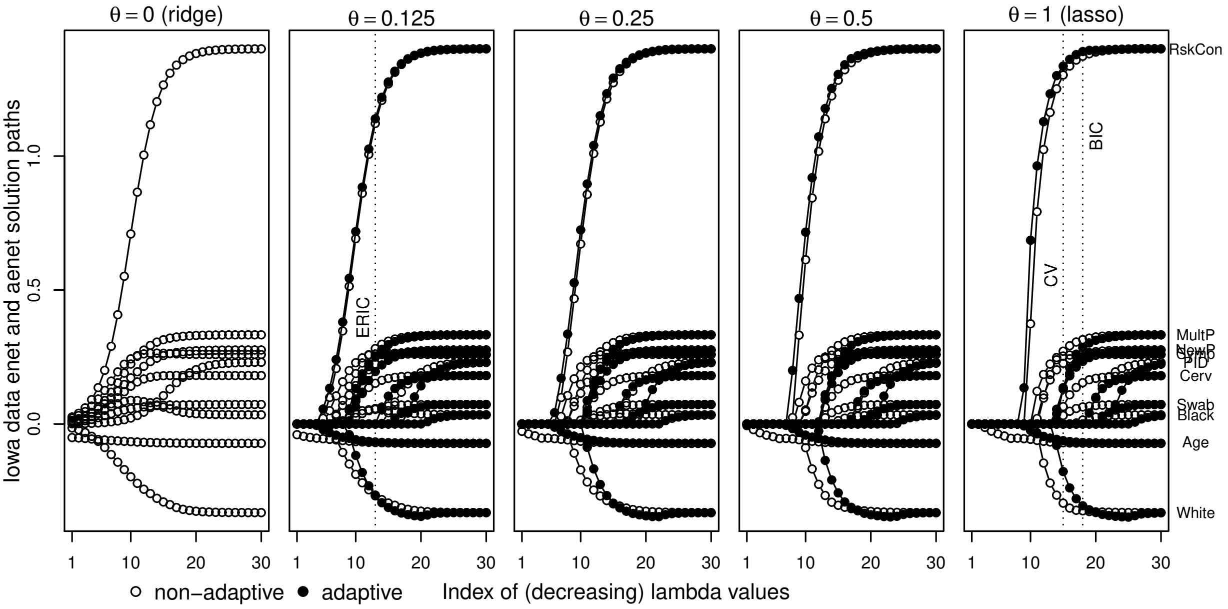
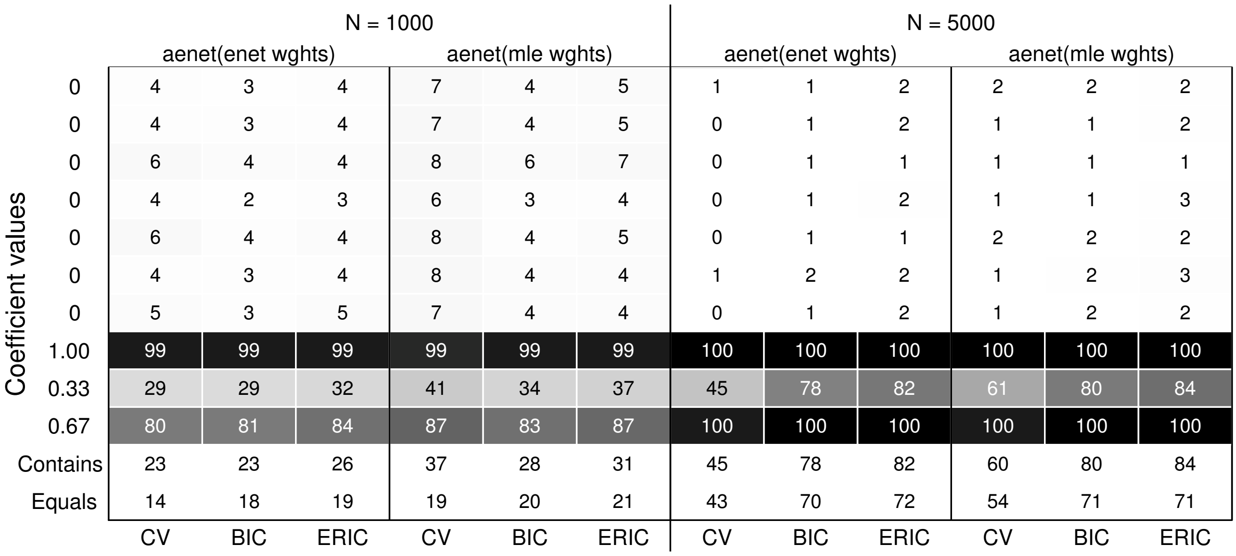


Figure 2. Elastic net and adaptive elastic net (with weights $\hat{\omega}^{\mathrm{mle}}$) solution paths for the Iowa Chlamydia data over 25 values of the tuning parameter λ for $\theta \in \{0, 1/8, 1/4, 1/2, 1\}$.



Coefficient values	N = 1000						N = 5000					
	aenet(enet wghts)			aenet(mle wghts)			aenet(enet wghts)			aenet(mle wghts)		
	CV	BIC	ERIC	CV	BIC	ERIC	CV	BIC	ERIC	CV	BIC	ERIC
0	4	3	4	7	4	5	1	1	2	2	2	2
0	4	3	4	7	4	5	0	1	2	1	1	2
0	6	4	4	8	6	7	0	1	1	1	1	1
0	4	2	3	6	3	4	0	1	2	1	1	3
0	6	4	4	8	4	5	0	1	1	2	2	2
0	4	3	4	8	4	4	1	2	2	1	2	3
0	5	3	5	7	4	4	0	1	2	1	2	2
1.00	99	99	99	99	99	99	100	100	100	100	100	100
0.33	29	29	32	41	34	37	45	78	82	61	80	84
0.67	80	81	84	87	83	87	100	100	100	100	100	100
Contains	23	23	26	37	28	31	45	78	82	60	80	84
Equals	14	18	19	19	20	21	43	70	72	54	71	71

Figure 3. Variable selection results under Model 1 under Dorfman testing from 500 simulations by the adaptive elastic net estimator under both choices of weights $\hat{\omega}^{\text{enet}}$ and $\hat{\omega}^{\text{mle}}$ under crossvalidation (CV), BIC, and ERIC choices of the tuning parameters λ and θ.

		aenet ($\hat{\omega}^{\text{mle}}$)			aenet ($\hat{\omega}^{\text{enet}}$)		
	mle	CV	BIC	ERIC	CV	BIC	ERIC
Intercept	*-0.93	-1.02	-0.88	-0.93	-0.89	-0.92	-1.00
Age	*-0.07	-0.06	-0.07	-0.07	-0.07	-0.07	-0.06
White	*-0.32	-0.18	-0.32	-0.27	-0.33	-0.32	-0.27
Black	0.05	.	0.02	0.05	.	.	.
NewP	*0.28	0.13	0.26	0.22	0.27	0.27	0.24
MultP	*0.33	0.14	0.31	0.28	0.33	0.33	0.29
RskCon	*1.40	1.34	1.37	1.12	1.40	1.38	1.26
Symp	*0.26	0.13	0.26	0.26	0.29	0.30	0.28
Swab	0.07	.	0.06	0.06	.	.	.
Cerv	0.18	.	0.15	0.15	.	.	.
PID	0.23	.	.	.	.	.	.
λ	.	8.89	2.49	20.77	14.54	2.16	10.83
θ	.	1	1	1/4	1/2	1/8	1/8

Table 1
The maximum likelihood estimator ($\hat{\alpha}^{\text{mle}}, \hat{\beta}^{\text{mle}}$) and the adaptive elastic net estimator ($\hat{\alpha}^{\text{aenet}}, \hat{\beta}^{\text{aenet}}$), with the weights $\hat{\omega}^{\text{mle}}$ and $\hat{\omega}^{\text{enet}}$, respectively, under the tuning parameter selections made by crossvalidation (CV), BIC, and ERIC. For the maximum likelihood estimator, '' denotes a p-value less than* 0.01 *when testing whether the coefficient is equal to zero. Tuning parameter selections are shown in the bottom two rows.*

				aenet ($\hat{\omega}^{\text{enet}}$)			aenet ($\hat{\omega}^{\text{mle}}$)		
model	procedure	oracle	mle	CV	BIC	ERIC	CV	BIC	ERIC
1	masterpool	1.47	**2.74**	**2.28**	2.30	2.39	**2.20**	2.30	2.38
	# assays: 1000	*(0.03)*	*(0.03)*	*(0.03)*	*(0.04)*	*(0.04)*	*(0.03)*	*(0.04)*	*(0.04)*
	Dorfman	0.96	**1.73**	1.34	**1.27**	1.38	1.32	**1.28**	1.38
	avg. # assays: 2212	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.02)*
	array	0.98	**1.67**	**1.38**	**1.31**	1.46	1.42	1.42	1.49
	avg. # assays: 2578	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*
	individual	1.02	**1.80**	1.41	**1.39**	1.53	**1.40**	1.43	1.54
	# assays: 5000	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*
2	masterpool	2.41	**4.15**	**3.33**	3.44	3.74	**3.30**	3.57	3.77
	# assays: 1000	*(0.03)*	*(0.03)*	*(0.04)*	*(0.05)*	*(0.05)*	*(0.03)*	*(0.05)*	*(0.05)*
	Dorfman	1.52	**2.59**	1.99	**1.72**	1.97	1.97	**1.90**	2.07
	avg. # assays: 2735	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*
	array	1.51	**2.59**	2.09	**1.81**	**1.95**	2.08	1.97	2.11
	avg. # assays: 2846	*(0.02)*	*(0.02)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*
	individual	1.56	**2.65**	2.04	**1.77**	2.01	2.02	**1.98**	2.13
	# assays: 5000	*(0.02)*	*(0.02)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*
3	masterpool	2.98	**5.32**	**4.43**	6.07	5.10	**4.40**	4.95	5.05
	# assays: 1000	*(0.03)*	*(0.03)*	*(0.04)*	*(0.16)*	*(0.08)*	*(0.04)*	*(0.08)*	*(0.07)*
	Dorfman	1.69	**2.85**	2.17	**1.91**	2.19	2.13	**2.12**	2.33
	avg. # assays: 2877	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*
	array	1.68	**2.85**	2.33	**2.04**	**2.24**	2.31	**2.24**	2.40
	avg. # assays: 2938	*(0.02)*	*(0.02)*	*(0.03)*	*(0.02)*	*(0.03)*	*(0.03)*	*(0.03)*	*(0.03)*
	individual	1.70	**2.91**	2.20	**1.91**	2.22	2.16	**2.14**	2.35
	# assays: 5000	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*	*(0.02)*

Table 2

Monte Carlo estimates over 500 simulation runs of $\mathbb{E}[\mathbb{E}\{\eta(\alpha_0 + X^T\beta_0) - \eta(\hat{\alpha} + X^T\hat{\beta})|\hat{\alpha}, \hat{\beta}\}^2]^{1/2} \times 100$ *at* $N = 5000$ *when the estimator* $(\hat{\alpha}, \hat{\beta}^T)^T$ *is the oracle estimator, the maximum likelihood estimator, and the adaptive elastic net estimators under crossvalidation (CV), BIC, and ERIC tuning parameter selection. Standard errors are given in parentheses.*