Supplementary material for "Combining Multiple Biomarker Models in Logistic Regression" by Yuan and Ghosh

1 Web Appendix A: Proof of Theorem 1

Let $D(p||q) = \int p(x) \log\{p(x)/q(x)\} d\mu(x)$ denote the Kullback-Leibler (K-L) divergence and $D_H^2(p,q) = \int(\sqrt{p(x)} - \sqrt{q(x)})^2 d\mu(x)$ denote the squared Hellinger distance between two densities p and q, respectively, with respect to some dominating measure μ. We have

$$\sum_{i=n/2+1}^{n} ED(P_f||P_{\tilde{f}_i}) = \sum_{i=n/2+1}^{n} E\int P_f(y_s)\log\frac{P_f(y_s)}{P_{\tilde{f}_i}(y_s)}v(dy)$$

$$= E\int P_f(y_s)\log\frac{\prod_{i=n/2+1}^{n} P_f(y_s)}{\prod_{i=n/2+1}^{n} P_{\tilde{f}_i}(y_s)}v(dy)$$

$$= E\int P_f(y_s)\log\frac{\prod_{i=n/2+1}^{n} P_f(y_s)}{\sum_j \lambda_j \prod_{i=n/2+1}^{n}\{(\hat{f}_j(x_s))^{y_s}(1-\hat{f}_j(x_s))^{1-y_s}\}}v(dy)$$

$$\leq E\int P_f(y_s)\log\frac{\prod_{i=n/2+1}^{n} P_f(y_s)}{\lambda_j \prod_{i=n/2+1}^{n}\{(\hat{f}_j(x_s))^{y_s}(1-\hat{f}_j(x_s))^{1-y_s}\}}v(dy)$$

$$\leq \log(1/\lambda_j) + \sum_{i=n/2+1}^{n} E\int P_f(y_s)\log\frac{P_f(y_s)}{P_{\hat{f}_j}(y_s)}v(dy)$$

for each $j \in \Gamma_s$. So we have

$$\sum_{i=n/2+1}^{n} ED(P_f||P_{\tilde{f}_i}) = \sum_{i=n/2+1}^{n} E\int P_f(y_s)\log\frac{P_f(y_s)}{P_{\tilde{f}_i}(y_s)}v(dy)$$

$$\leq \inf_{j\in\Gamma_s}\{\log(1/\lambda_j) + \sum_{i=n/2+1}^{n} E\int P_f(y_s)\log\frac{P_f(y_s)}{P_{\hat{f}_j}(y_s)}v(dy)\}$$

Since

$$\int P_f(y_s)\log\frac{P_f(y_s)}{P_{\hat{f}_j}(y_s)}v(dy) = f(x_s)\log\frac{f(x_s)}{\hat{f}_j(x_s)} + (1-f(x_s))\log\frac{(1-f(x_s))}{(1-\hat{f}_j(x_s))}$$

$$\leq \frac{(f-\hat{f}_j)^2}{\hat{f}_j} + \frac{(f-\hat{f}_j)^2}{1-\hat{f}_j} \leq \frac{2}{A_j^2}(f-\hat{f}_j)^2$$

where we assume that $A_j \leq \hat{f}_j \leq 1 - A_j$ for all x. Therefore

$$\sum_{i=n/2+1}^{n} ED(P_f||P_{\tilde{f}_i}) \leq \inf_{j \in \Gamma_s} \{\log(1/\lambda_j) + \frac{2}{A_j^2} \sum_{i=n/2+1}^{n} E(f - \hat{f}_j)^2\}$$

Since the K-L divergence is always lower bounded by the squared Hellinger distance, we have

$$D(P_f||P_{\tilde{f}_i}) \geq D_H^2(P_f, P_{\tilde{f}_i}) = [\sqrt{f} - \sqrt{\tilde{f}_i}]^2 + [\sqrt{(1-f)} - \sqrt{(1-\tilde{f}_i)}]^2$$

$$\geq \frac{1}{4}((f - \tilde{f}_i)^2 + (f - \tilde{f}_i)^2) = \frac{1}{2}(f - \tilde{f}_i)^2$$

By taking the expectation on both sides of the above inequality, for any $j \in \Gamma_s$ if $\lambda_j = 1/K_s$, we have

$$\sum_{i=n/2+1}^{n} ||f - \tilde{f}_i||^2 \leq 2 \sum_{i=n/2+1}^{n} ED(P_f||P_{\tilde{f}_i})$$

$$\leq 2 \inf_{j \in \Gamma_s} \{\log(1/\lambda_j) + \frac{2}{A_j^2} \sum_{i=n/2+1}^{n} E(f - \hat{f}_j)^2\}$$

$$= 2 \inf_{j \in \Gamma_s} \{\log(1/\lambda_j) + \frac{2}{A_j^2} \sum_{i=n/2+1}^{n} ||f - \hat{f}_j||^2\}$$

$$\leq 2\{\log(K_s) + \frac{2}{A_j^2} \sum_{i=n/2+1}^{n} ||f - \hat{f}_j||^2\}.$$

Then by the definition of $\hat{f}_n^*$, for any $j \in \Gamma_s$, we have

$$||f - \hat{f}_n^*||^2 = \frac{1}{n/2} \sum_{i=n/2+1}^{n} ||f - \tilde{f}_i||^2$$

$$\leq \frac{2}{n/2}\{\log(K_s) + \frac{2}{A_j^2} \sum_{i=n/2+1}^{n} ||f - \hat{f}_j||^2\}$$

$$\leq 2\{\frac{\log(K_s)}{n/2} + \frac{2}{A_j^2} \frac{1}{n/2} \sum_{i=n/2+1}^{n} ||f - \hat{f}_j||^2\}$$

$$= 2\{\frac{\log(K_s)}{n/2} + \frac{2}{A_j^2} ||f - \hat{f}_j||^2\}.$$

Since $\hat{f}_n^*$ is a convex combination of the original estimators, under Condition 2, we have $||f - \hat{f}_n^*||^2 \leq \tau$ when j is not in Γ_s. Therefore it follows that

$$||f - \hat{f}_n^*||^2 \leq \tau(1 - P(j \in \Gamma_s)) + 2\{\frac{\log(K_s)}{n/2} + \frac{2}{A_j^2}||f - \hat{f}_j||^2\}P(j \in \Gamma_s).$$

The conclusion then follows. This completes the proof of Theorem 1.

2 Web Appendix B: Definition and Estimation of Generalized Degrees of Freedom (GDF)

Consider a normal response vector $Y \equiv (y_1, \ldots, y_n)^{'}$, where

$$Y \sim N(\mu, \sigma^2 I).$$

where $\mu \equiv (\mu_1, ..., \mu_n)^{'}$ is a mean vector. Define a modelling procedure M to be a mapping from R^n to R^n that produces a set of fitted values $\hat{\mu} = (\hat{\mu}_1, ..., \hat{\mu}_n)^{'}$ from Y. That is, $M : Y \to \hat{\mu}$. The GDF for a modelling procedure M (Ye, 1998) is given by $D(M) = \sum_{i=1}^{n} h_i^M(\mu)$, where

$$\begin{aligned} h_i^M(\mu) &= \frac{dE_\mu[\hat{\mu}_i(Y)]}{d\mu_i} = \lim_{\delta \to 0} E_\mu[\frac{\hat{\mu}_i(Y + \delta e_i) - \hat{\mu}_i(Y)}{\delta}] \\ &= \frac{1}{\sigma^2} E[\hat{\mu}_i(Y)(y_i - \mu_i)] = \frac{1}{\sigma^2} cov(\hat{\mu}_i(Y), y_i - \mu_i), \end{aligned}$$

where e_i is the ith column of the n by n identity matrix. The GDF is defined to be the sum of the average sensitivities of the fitted value $\hat{\mu}_i(Y)$ to a small change in y_i. Thus it measures the flexibility of the modelling procedure M. If M is highly flexible, then the fitted values tend to be close to the observed values. Thus the sensitivity of the fitted values to the observed values would be high, and the GDF would be large.

Shen et al. (2004) showed that the GDF for general exponential family is $D(M) = \sum_{i=1}^{n} h_i^M(\mu)$ with $h_i^M(\mu) = E\phi(\hat{\mu}_i)(Y_i - \mu_i)$, where the density of the exponential family is defined as

$$p(y_i|\mu_i) = \exp(\phi(\mu_i)y_i + \alpha(\mu_i) + m(y_i))$$

with mean $\mu_i = EY_i$ and variance $var(Y_i) = 1/\phi^{'}(\mu_i)$. They proposed an estimator of GDF using the idea of data perturbation. We will describe the details for calculating GDF for logistic regression models.

For a logistic model, Y_i is distributed according to the binomial distribution $B(1, p(X_i))$, $\mu_i = E(Y_i) = p(X_i)$ and $\phi(\mu_i) = log(p(X_i)/(1 - p(X_i)))$. Define $\tilde{Y}_i$ as an independent binary random variable distributed according to $B(1, \tilde{p}_i)$ with a pre-specified $\tilde{p}_i$ $(0 < \tilde{p}_i < 1)$ or estimated $\tilde{p}_i = \hat{p}(X_i)$. Then define Y_i^* as $(1-\tau)Y_i + \tau\tilde{Y}_i, i = 1, \ldots, n$, where $0 \leq \tau \leq 1$. We can see than the conditional distribution of Y_i^* given Y_i has the same support range. In logistic regression, the same estimation method based on the iterative reweighted least squares method is directly applicable to $Y^* = (Y_1^*, \ldots, Y_n^*)'$ even if each Y_i^* may no longer be binary. This can be thought of as embedding a specific model into a more general class of models defined by the exponential family distribution. In their Theorem 1, Shen et al. (2004)in show that an estimator of GDF is given by

$$\widehat{GDF}(M) = \frac{1}{\tau^2}\sum_{i=1}^{n} cov^*(\phi(\hat{\mu}_i(Y^*)), Y_i^*),$$

where cov^* is the conditional covariance given Y. In general, $\widehat{GDF}(M)$ may be computed using a Monte Carlo numerical approximation. The algorithm for computing $\widehat{GDF}(M)$ works as follows:

1. First, we sample $\eta_i^{(j)}$ independently from the distribution of Y_i^* as described earlier for $i = 1, \ldots, n$ and $j = 1, \ldots, T$.

2. Second, we compute $\{\hat{\mu}(\eta_i^{(j)}) : j = 1, \ldots, T\}$. For each i, compute

$$\hat{b}_i = \frac{1}{T-1}\sum_{j=1}^{T}(\phi(\hat{\mu}(\eta_i^{(j)})) - \bar{\phi}_i)(\eta_i^{(j)} - \bar{\eta}_i),$$

   where $\bar{\eta}_i = \frac{1}{T}\sum_{j=1}^{T}\eta_i^{(j)}$ and $\bar{\phi}_i = \frac{1}{T}\sum_{j=1}^{T}\phi(\hat{\mu}(\eta_i^{(j)}))$ are the Monte Carlo means.

3. Then $\widehat{GDF}(M)$ is approximated by $\frac{1}{\tau^2}\sum_{i=1}^{n}\hat{b}_i$. Here T is chosen to be sufficiently large to ensure approximation precision. We chose T to be n and τ to be 0.5.

3 Web Appendix C: Simulation studies

Case 2: We use the following true model for relating biomarkers with disease status:

$$\text{logit}P(D=1) = 1.0 + 1.1X_1 + 1.2X_2$$

This model only includes two biomarker predictors with relatively large coefficients. The uncertainty of model selection procedures in this case is relatively small compared to that in Case 1. The results are shown in Tables 1 and 2. We observe that all combining methods perform similarly to the AIC-selected model due to small uncertainty in this case. For the sample size of 100, the ARMS methods with the GDF, AIC and Bernoulli weights have slightly smaller prediction risks and slightly higher AUC values than the AIC-selected model. For the sample size of 400, the ARMS methods with the GDF, AIC and Bernoulli weights have slightly larger prediction risk and smaller AUC values than the AIC-selected model. For both sample sizes, BMA performs slightly better than ARMS. Hence, our ARMS-GDF method performs significantly better than the BMA method in Case 1, while it performs just slightly worse than the BMA method when in Case 2.

4 Web Appendix D: Adaptive model screening algorithm

In order to apply our combining algorithm on higher-dimensional data, we propose a new screening method for our ARMS algorithm using the adaptive penalty for model selection (Shen et al., 2004). We replace step 2 of the ARMS algorithm by the following:

- Step 2. Estimate β_k by $\hat{\beta}_k$ using maximum likelihood for each candidate logistic model M_k based on the first half of the data $Q^{(1)}$ and let $\hat{p}_k^{(Q^{(1)})}(\mathbf{x}_i;\hat{\beta}_k)$ be the estimate of $p_k(\mathbf{x}_i;\beta_k) = e^{\mathbf{x_{ki}}\beta}/(1+e^{\mathbf{x_{ki}}\beta})$, where $\mathbf{x_k}$ is a subset of $\mathbf{x}$. Compute the adaptive model selection criterion

$$-\sum_{i=1}^{i=n/2} \log[\{\hat{p}_k^{(Q^{(1)})}(\mathbf{x}_i;\hat{\beta}_k)\}^{D_i}\{1-\hat{p}_k^{(Q^{(1)})}(\mathbf{x}_i;\hat{\beta}_k)\}^{1-D_i}] + \lambda|M_k|,$$

Table 1: ARMS simulation results of Case 2 with $n = 100$.

Method	L_2 risk	L_1 risk	EP	AUC
ARMS-LIKELI	0.041	0.093	0.398	0.821
	(0.002)	(0.003)	(0.007)	(0.010)
ARMS-AIC	0.041	0.093	0.396	0.823
	(0.002)	(0.003)	(0.007)	(0.010)
ARMS-GDF	0.040	0.091	0.395	0.826
	(0.002)	(0.003)	(0.007)	(0.010)
ARMS-APE	0.042	0.094	0.402	0.811
	(0.002)	(0.004)	(0.008)	(0.011)
ARMS-RESID	0.042	0.095	0.405	0.813
	(0.002)	(0.004)	(0.008)	(0.011)
BMA	0.039	0.089	0.391	0.831
	(0.002)	(0.003)	(0.007)	(0.010)
AIC	0.042	0.094	0.404	0.809
	(0.002)	(0.004)	(0.008)	(0.011)
Full	0.063	0.140	0.436	0.728
	(0.003)	(0.006)	(0.011)	(0.014)
True	0.033	0.075	0.366	0.899

Note: See Note to Table 1 in the paper.

Table 2: ARMS simulation results of Case 2 with $n = 400$.

Method	L_2 risk	L_1 risk	EP	AUC
ARMS-LIKELI	0.034	0.067	0.389	0.841
	(0.001)	(0.002)	(0.004)	(0.007)
ARMS-AIC	0.034	0.067	0.389	0.840
	(0.001)	(0.002)	(0.004)	(0.007)
ARMS-GDF	0.034	0.066	0.388	0.842
	(0.001)	(0.002)	(0.004)	(0.007)
ARMS-APE	0.035	0.068	0.391	0.838
	(0.001)	(0.002)	(0.005)	(0.007)
ARMS-RESID	0.035	0.068	0.391	0.839
	(0.001)	(0.002)	(0.005)	(0.007)
BMA	0.0032	0.063	0.381	0.853
	(0.001)	(0.002)	(0.004)	(0.007)
AIC	0.033	0.065	0.385	0.847
	(0.001)	(0.002)	(0.005)	(0.007)
Full	0.048	0.093	0.413	0.756
	(0.002)	(0.004)	(0.007)	(0.009)
True	0.027	0.054	0.355	0.911

Note: See Note to Table 1 in the paper.

indexed by the penalty parameter λ, where λ is chosen from 0 to 10 by 0.1 increment. $|M|$ is the size of the model M, i.e. the number of parameters. We find the optimal model M_λ for each λ using the backward stepwise searching proposed by Shen et al. (2004). Then the 100 selected optimal models under the different penalty parameters are considered as candidates for combining. Then we follow the same steps in the original algorithm.

We call this new ARMS algorithm the ARMS-adaptive algorithm. Now we want to compare the new algorithm with the old one using the prostate cancer data. We did exactly the same for the old ARMS, while we enlarge our model space by including all possible pairwise interactions in the model for the new ARMS-adaptive algorithm. This results in 2^{36} subset models for the new model space. The results are in Table 3.

The results show that the new ARMS-adaptive algorithm has some further improvement over the old ARMS algorithm. There are several possible reasons for this. First, the ARMS-adaptive algorithm can deal with a much larger model space and higher data dimension than the old ARMS algorithm. Second, the ARMS-adaptive algorithm uses an adaptive penalty parameter for screening which is less sensitive to the size of underlying true model than the AIC-type screening in the old ARMS algorithm. However, we note that the ARMS-adaptive method does not search the complete model space exhaustively, as did our original algorithm. We believe that in most high-dimensional data cases, the gain from larger model space and adaptive penalty parameter will offset the loss due to incomplete stepwise searching. However, further work in this area is needed.

Shen, X., Huang, H. and Ye, J. (2004). Adaptive model selection and assessment for exponential family distributions. *Technometrics* **46**, 306–317.

Table 3: Comparison of ARMS using AIC screening with ARMS-adaptive algorithm under GDF weights (1000 permutations).

Method	L_2 risk	L_1 risk	EP	AUC
AIC	0.089 (0.003)	0.179 (0.05)	0.405 (0.007)	0.76 (0.01)
ARMS	0.078 (0.002)	0.159 (0.004)	0.373 (0.006)	0.82 (0.01)
ARMS-adaptive	0.074 (0.002)	0.148 (0.004)	0.365 (0.006)	0.84 (0.01)