Statistical Methods in Medical Research
27(X):1–??
©The Author(s) 2018
Reprints and permission:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/ToBeAssigned
www.sagepub.com/
SAGE

Supplementary material to High dimensional prediction of binary outcomes in the presence of between-study heterogeneity

Chamberlain Mbah[1] Jan De Neve [2] Olivier Thas [34]

[1] Department of Radiotherapy and Experimental Cancer Research, Ghent University, Ghent, Belgium
[2] Department of Data Analysis, Faculty of Psychology and Educational Sciences, Ghent University, Ghent, Belgium
[3] Center for Statistics, Hasselt University, Diepenbeek, Belgium
[4] National Institute for Applied Statistics Research Australia, University of Wollongong, South Wales, Australia

Corresponding author:
Chamberlain Mbah
Department of Radiotherapy and Experimental Cancer Research, Ghent University, Ghent, Belgium
C. Heymanslaan 10, Radiotherapiepark,9000 Ghent, Belgium
Email: chamberlain.mbah@ugent.be

1 Simulations with imbalanced outcomes

We extend the simulations study in the main paper with an imbalanced outcome ($n_+ = 12$, $n_- = 48$) with $\alpha_0 = 0.025$. We also introduce the area under the precision recall curve (AUC-PR) as a performance metric as it is better suited when the outcome is imbalanced. The results are displayed in Table 1.

For $J = 4$ the performance of the different estimators in prediction have dropped compared to the balanced case in the main paper. However, the relative performance of the different estimators is similar to that of the main paper. That is, EB still has the best performance among the other estimators. Ridge and and the shrunken centroids estimators generally perform worse than their empirical Bayes counterparts: EBComBat, EBMerge and EBAvg. The MLEs $\hat{\Delta}_i$ and $\hat{\beta}_i$ have the worst performances amongst all estimators.

For $J = 15$ the performance of the different estimators in prediction improves greatly compared to $J = 4$, EB is still top in performance, EBComBat and the MLE $\hat{\beta}_i$ have performances that are close to that of EB. The MLE $\hat{\Delta}_i$ has the worst performance amongst all other estimators.

2 Simulations on correlations between predictors

In this simulation study, we assess the effect of correlation between predictors on the prediction performance of the empirical Bayes estimator, EB. We start by simulating data independently and move on to incorporate a range of correlations – from weak to strong. The prediction performance of the rule in (6) of the main paper with EB estimators is assessed over the span of correlations. The goal is to examine how poorly the rule in (6) (of the main paper) performs with increasing correlation between predictors.

Table 1. For $J = 4$ and $J = 15$ studies, performance metrics, standard errors in parenthesis and the cardinality of $\mathcal{S}_{\text{final}}$. Estimators are ranked according to their misclassification error rate estimate $\hat{\alpha}_{\text{CV}}(\mathcal{S}_{\text{final}})$. The performance of the true parameter values β_i and Δ_i, are shown for benchmarking. The performance of the maximum likelihood estimators (MLE) of Δ_i and β_i, $\hat{\Delta}_i$ and $\hat{\beta}_i$, respectively, are also presented. The imbalanced in the outcome is 20% vs 80%. Cross-validation estimates of the area under the precision recall curve (AUC-PR$_{\text{CV}}$) and the area under the receiver characteristic curve (AUC-ROC$_{\text{CV}}$) are shown.

$J = 4$

	$\hat{\alpha}_{\text{CV}}$	ExpLoss$_{\text{CV}}$	AUC-ROC$_{\text{CV}}$	AUC-PR$_{\text{CV}}$	Cardinality of $\mathcal{S}_{\text{final}}$
true β_i	0.015 (0.03)	0.583 (0.08)	0.999 (0.00)	0.913 (0.00)	41
true Δ_i	0.083 (0.02)	0.594 (0.05)	0.965 (0.00)	0.826 (0.00)	14
EB	0.293 (0.08)	0.621 (0.11)	0.790 (0.09)	0.467 (0.01)	12
EBMerge	0.307 (0.07)	0.618 (0.10)	0.777 (0.07)	0.451 (0.09)	34
EBAvg	0.332 (0.09)	0.618 (0.05)	0.752 (0.07)	0.423 (0.08)	22
EBComBat	0.377 (0.06)	0.622 (0.43)	0.666 (0.08)	0.337 (0.12)	8
Ridge	0.378 (0.08)	0.625 (0.42)	0.672 (0.07)	0.362 (0.41)	8
Shrunken centroids	0.380 (0.05)	0.623 (0.18)	0.667 (0.09)	0.334 (0.22)	13
MLE $\hat{\Delta}_i$	0.443 (0.10)	0.624 (0.38)	0.581 (0.21)	0.146 (0.40)	1
MLE $\hat{\beta}_i$	0.457 (0.12)	0.625 (0.23)	0.572 (0.11)	0.202 (0.20)	4

$J = 15$

	$\hat{\alpha}_{\text{CV}}$	ExpLoss$_{\text{CV}}$	AUC-ROC$_{\text{CV}}$	AUC-PR$_{\text{CV}}$	Cardinality of $\mathcal{S}_{\text{final}}$
true β_i	0.010 (0.03)	0.584 (0.08)	1.000 (0.00)	0.915 (0.02)	41
EB	0.037 (0.08)	0.585 (0.08)	0.992 (0.05)	0.896 (0.09)	26
EBComBat	0.042 (0.06)	0.586 (0.09)	0.995 (0.09)	0.899 (0.07)	53
MLE $\hat{\beta}_i$	0.045 (0.03)	0.589 (0.12)	0.992 (0.09)	0.888 (0.08)	32
true Δ_i	0.068 (0.04)	0.588 (0.04)	0.985 (0.00)	0.872 (0.01)	14
EBMerge	0.080 (0.05)	0.591 (0.09)	0.980 (0.09)	0.865 (0.09)	50
EBAvg	0.083 (0.06)	0.594 (0.12)	0.980 (0.11)	0.856 (0.26)	14
Shrunken centroids	0.090 (0.03)	0.594 (0.11)	0.970 (0.14)	0.832(0.14)	26
Ridge	0.103 (0.07)	0.617 (0.11)	0.943 (0.16)	0.774 (0.13)	58
MLE $\hat{\Delta}_i$	0.123 (0.08)	0.601 (0.32)	0.940 (0.28)	0.766 (0.18)	12

For such comparisons to make sense, it is important that the effect sizes of the predictors remain the same throughout the range of correlations. This way, only the effect of the correlation between predictors is investigated.

In the main text we gave two parametric forms of the misclassification error rate α. An α without and with correlation correction, $\hat{\alpha}$ and $\hat{\alpha}_{\text{cor}}$ respectively. As a secondary goal, we will compare the performance of models selected with $\hat{\alpha}$ and $\hat{\alpha}_{\text{cor}}$.

Finally, we know that $\hat{\alpha}$ reduces when more predictors are added to the prediction rule and eventually leads to too optimistic conclusions about model performance. With cross validation, a close to honest estimate of model performance is possible. In the main text, $\hat{\alpha}_{\text{CV}}$ was introduced as a cross validation estimate of $\hat{\alpha}$ an we compare $\hat{\alpha}_{\text{cor}}$ and $\hat{\alpha}_{\text{CV}}$.

2.1 Generating data

We replicate 1000 Monte Carlo runs and average the results over these runs. In each Monte Carlo run, data for $n = 60$ subjects is generated within each of the J studies, with $n_{-j} = 30$ stable $(Y = -1)$ and $n_{+j} = 30$ reject $(Y = +1)$ subjects, $j = 1, \ldots, J$. The number of studies is set to $J = 4$ and the number of genes to $N = 10000$. We denote the standard normal distribution with Φ and generate the effect sizes as

$$\begin{aligned} \Delta_i &\sim \Phi^{-1}\left(\frac{i}{301}\right), \quad i = 1 \ldots, 300 \text{ and} \\ \Delta_i &= 0, \quad i = 301, \ldots, N. \end{aligned}$$

In this way, only the first 300 genes are predictive for the outcome. The heterogeneity parameters are generated as

$$\tau_i^2 \sim F_{\chi_1}^{-1}\left(\frac{i}{N+1}\right) + 1, \quad i = 1, \ldots, N,$$

where $F_{\chi_1^2}^{-1}(x)$ is the quantile function of a chi-squared distribution with one degree of freedom. Note that $\tau_i^2 \geq 1$. We begin with simulating the predictors independently and include correlations in the next step, i.e.,

$$\delta_{ij} \sim \mathcal{N}\left(\Delta_i, \sigma_i^2\right) \text{ and } \sigma_i^2 = 4(\tau_i^2 - 1)$$

with

$$X'_{ij}|Y_j \sim \mathcal{N}\left(\mu_{ij} + \lambda_{ij} Y_j \delta_{ij}/2, \lambda_{ij}^2\right),$$

we then let

$$U'_{ij} = \frac{X'_{ij} - \mu_{ij}}{\lambda_{ij}},$$

so that

$$U'_{ij} \mid Y_j, \Delta_i \sim N(Y_j\Delta_i/2, \tau_i^2).$$

Now we add the correlation as the parameter ρ_i, for the ith predictor with $\rho_i \in [0, 1)$, and generate the correlated predictors as

$$\begin{aligned} X_{ij} &= \sqrt{\rho_i}\epsilon + \sqrt{1-\rho_i}U'_{ij}, \\ \text{so that } X_{ij}|Y &\sim \mathcal{N}\left(\sqrt{1-\rho_i}Y\Delta_i/2, \rho_i + (1-\rho_i)\tau_i^2)\right). \end{aligned} \tag{1}$$

Here, ϵ is a standard Gaussian random variable and we assume equal variances and covariances in the two groups ($Y = \pm 1$). Equation (1) ensures that the covariance between predictors X_i and X_i' is $\sqrt{\rho_i\rho_{i'}}$. This follows from,

$$\begin{aligned} \text{Cov}\left(X_{ij}, X_{i'j}\right) &= \text{E}\left\{X_iX_{i'}\right\} - \text{E}\left\{X_{ij}\right\}\text{E}\left\{X_{i'j}\right\} \\ &= \text{E}\left\{\epsilon^2\sqrt{\rho_i\rho_{i'}} + \rho_i\epsilon\sqrt{1-\rho_i}U_{ij}\right\} \\ &+ \text{E}\left\{\rho_{i'}\epsilon\sqrt{1-\rho_{i'}}U'_{i'j} + \sqrt{(1-\rho_i)(1-\rho_{i'})}U_{ij}U'_{i'j}\right\} \\ &- \text{E}\left\{X_{ij}\right\}\text{E}\left\{X_{i'j}\right\} \\ \\ &= \text{E}\left\{\epsilon^2\sqrt{\rho_i\rho_{i'}}\right\} + \text{E}\left\{\sqrt{(1-\rho_i)(1-\rho_{i'})}U_{ij}U'_{i'j}\right\} \\ &- \text{E}\left\{\sqrt{(1-\rho_i)(1-\rho_{i'})}U_{ij}U'_{i'j}\right\} \\ &= \sqrt{\rho_i\rho_{i'}}, \text{ since } \text{E}\left\{\epsilon^2\right\} = 1. \end{aligned}$$

Equation (1) also makes sure that the effect size of each predictor remains the same after adding correlations between the predictors. To see this, note that the standardized predictor X_{ij} denoted as U_{ij} takes the form

$$U_{ij} = \frac{X_{ij}}{\sqrt{1-\rho_i}} \sim \mathcal{N}\left(Y\frac{\Delta_i}{2}, \frac{\rho_i}{1-\rho_i} + \tau_i^2\right),$$

with

$$\mathrm{Cov}\left(U_{ij}, U_{i'j}\right) = \sqrt{\frac{\rho_i \rho_{i'}}{(1-\rho_i)(1-\rho_{i'})}}.$$

So that

$$\Delta_i = \mathrm{E}\left\{U_{ij} \mid Y = 1\right\} - \mathrm{E}\left\{U_{ij} \mid Y = -1\right\},$$

and Δ_i is interpreted as the effect size of the predictor U_{ij}.

2.2 Assessments

We compute the empirical Bayes estimate of β_i, EB, from the correlated training dataset. EB is used in the rule (6) (of the main text) to make predictions on a correlated test dataset (generated in same way as the training), and its performance is assessed with $\hat{\alpha}_{\mathrm{CV}}$, $\mathrm{ExpLoss}_{\mathrm{CV}}$, $\mathrm{Brier}_{\mathrm{CV}}$ and $\mathrm{AUC}_{\mathrm{CV}}$. For benchmarking, we also assess the performance of the true values of β_i in prediction, when used in the prediction rule in (6) of the main paper. This is done for each Monte Carlo run. All correlations between predictors are set to ρ, ρ taking values $0.1, 0.3, 0.6$ and 0.9. We will use the two model selection procedures discussed in the main text: 1) the model selection approach in Section 4.2 with $\alpha_0 = 0.025$, referred to as Method 1 here. This model selection method uses α estimates without correlation correction for model selection. 2) The model selection with estimates of α_{cor} as described in Section 6 of the main paper, this method is referred to as Method 2. Here, the model selection is with respect α, corrected for the correlations between predictors. Since we are comparing two model selection criteria, we make a distinction between the metrics used for assessing these model selection criteria. Using Method 1 for model selection, we arrive at the optimal set of predictors $\mathcal{S}_{\mathrm{final}}$, and it is denoted here as $\mathcal{S}_{\mathrm{final1}}$. We add the subscript 1, to all the metrics used for assessing variable selection with Method 1 ($\hat{\alpha}_{\mathrm{CV1}}$, $\mathrm{ExpLoss}_{\mathrm{CV1}}$, $\mathrm{Brier}_{\mathrm{CV1}}$ and $\mathrm{AUC}_{\mathrm{CV1}}$). Using Method 2, $\hat{\alpha}_{\mathrm{cor}}$ attains a minimum after a certain number (ordered by their magnitude) of predictors have been added to the

Table 2. Empirical Bayes (EB) and the true β performance in classification when used in rule (6) (of the main paper), at different values of the correlation, ρ. Their performance is assessed with α from cross validation (CV), $\hat{\alpha}_{\mathrm{CV}}$, CV exponential loss, $\mathrm{ExpLoss}_{\mathrm{CV}}$, CV Brier score, $\mathrm{Brier}_{\mathrm{CV}}$, and the area under the receiver characteristic curve, $\mathrm{AUC}_{\mathrm{CV}}$, from CV. The subscripts 1 and 2 added to the model assessment metrics, refer to the method used for model selection (Method 1 and Method 2). The cardinality of the optimal set of predictors using the two model selection methods are also shown along side the estimate of $\hat{\alpha}_{\mathrm{cor}}$. Model performance drops with increasing ρ.

Estimators	$\hat{\alpha}_{\mathrm{CV1}}$	$\hat{\alpha}_{\mathrm{CV2}}$	$\mathrm{ExpLoss}_{\mathrm{CV1}}$	$\mathrm{ExpLoss}_{\mathrm{CV2}}$	$\mathrm{Brier}_{\mathrm{CV1}}$	$\mathrm{Brier}_{\mathrm{CV2}}$	$\mathrm{AUC}_{\mathrm{CV1}}$	$\mathrm{AUC}_{\mathrm{CV2}}$	Cardinality $\mathcal{S}_{\mathrm{final1}}$	Cardinality $\mathcal{S}_{\mathrm{final2}}$	$\hat{\alpha}_{\mathrm{cor}}$
						$\rho = 0.1$					
true β	0.026	0.039	0.093	0.134	0.020	0.029	0.997	0.993	27	19	0.058
EB	0.175	0.320	0.561	0.728	0.126	0.208	0.901	0.741	18	12	0.228
						$\rho = 0.3$					
true β	0.032	0.256	0.113	0.621	0.024	0.172	0.996	0.824	27	15	0.225
EB	0.219	0.329	0.596	0.701	0.153	0.211	0.856	0.732	11	9	0.215
						$\rho = 0.6$					
true β	0.067	0.351	0.313	0.870	0.051	0.228	0.983	0.706	27	12	0.247
EB	0.261	0.376	0.757	0.801	0.177	0.234	0.812	0.667	9	7	0.264
						$\rho = 0.9$					
true β	0.232	0.429	2.801	2.079	0.206	0.329	0.850	0.598	27	3	0.272
EB	0.448	0.475	0.924	0.715	0.245	0.251	0.573	0.537	8	1	0.426

prediction rule; we refer to this set as $\mathcal{S}_{\mathrm{final2}}$. The cross validation estimates of α, the exponential loss, the Brier score and the AUC computed using the predictors in $\mathcal{S}_{\mathrm{final2}}$, are referred to as $\hat{\alpha}_{\mathrm{CV2}}$, $\mathrm{ExpLoss}_{\mathrm{CV2}}$, $\mathrm{Brier}_{\mathrm{CV2}}$ and $\mathrm{AUC}_{\mathrm{CV2}}$.

2.3 Results, remarks and conclusions

The following remarks and conclusions are drawn from this simulation study.

In Table 2, we present the results of our simulations, for β and EB, with increasing values of ρ. Model performance deteriorates with increasing ρ, this holds for β and EB. However, the rate at which the models deteriorate is rather slow. Only very high correlations, $\rho = 0.9$, completely break the performance of the EB estimator.

Using Method 2 for variable selection, results in too few predictors, the cardinality of $\mathcal{S}_{\mathrm{fina12}}$ on Table 2. For this reason, $\hat{\alpha}_{\mathrm{CV2}}$, $\mathrm{ExpLoss}_{\mathrm{CV2}}$, $\mathrm{Brier}_{\mathrm{CV2}}$ and $\mathrm{AUC}_{\mathrm{CV2}}$ are all worse than their counterparts computed via Method 1. We may conclude that, model selection with Method 2 is too conservative and results in worse performance.

Note that $\mathcal{S}_{\text{final2}}$ is the index set of predictors that yielded $\hat{\alpha}_{\text{cor}}$, and $\hat{\alpha}_{\text{CV2}}$ is a near honest estimate of α for the set $\mathcal{S}_{\text{final2}}$. We see that $\hat{\alpha}_{\text{cor}}$ is overly optimistic about α.