Handling missing probe-sets in a linear gene classifier

Rowan Kuiper, Erasmus MC Cancer Institute, Rotterdam, the Netherlands.

November 11 2016

In multiple myeloma (a type of cancer) linear models (e.g. EMC92) can be used to estimate a patients' prognosis. In case the model outcome exceeds a specific value t (i.e. a dichotomizing threshold) the subject is classified as high-risk. These linear models consist of a number of covariates, each contributing to the model outcome. In the case of EMC92 or UAMS70, these covariates are probe sets. This vignette deals with the situation that not all covariates are available for generating the model outcome in independent data. The method is based on redistributing the weights of the discarded covariates over the remaining covariates based on the covariance structure in the training data of that model, i.e. a **reweighted model**.

1 Effects of discarding variables from a model

Consider a linear model that enables the identification of patients with worse survival. Let the model be $\hat{\boldsymbol{Y}} = \boldsymbol{X}\boldsymbol{\beta}$ with outcome $\boldsymbol{Y} \in \mathbb{R}$ being a function of the vector of probe-set weights $\boldsymbol{\beta}$ and the $n \times m$ gene expression matrix $\boldsymbol{X}$ containing n patients and m probe-sets. Now consider the situation of interest, where only a subset of covariates in $\boldsymbol{X}$ indexed by $j \in \{1..m\}$ is non-missing. Often, missing covariates are simply discarded from the prediction such that the reduced model is $\hat{\boldsymbol{Y}}^R = \boldsymbol{X}_{.j}\boldsymbol{\beta}_j$.

To illustrate the effect of discarding covariates on the outcome consider a linear model with two variables: $\hat{y} = x_1\beta_1 + x_2\beta_2$ which has an expected value of $E[\hat{y}] = E[x_1]\beta_1 + E[x_2]\beta_2$. By disregarding covariate x_2 we are left with the reduced model $\hat{y}^R = x_1\beta_1$ which has a biased outcome compared to the complete model if $E[x_2] \neq 0$ - with an expected value of $E[\hat{y}] = E[x_1]\beta_1$. The same is true for the variance. To illustrated this, we further simplify the example by mean centering each $\boldsymbol{x}$ such that $E[\boldsymbol{x}] = 0$. This will result in a variance for the complete model of $Var[\hat{\boldsymbol{y}}] = \boldsymbol{\beta}\boldsymbol{\Sigma}\boldsymbol{\beta}' = \beta_1^2 + \beta_2^2 + \beta_1\beta_2 2Cov(x_1, x_2)$ for $\boldsymbol{\Sigma} = Cov(x_1, x_2)$. By disregarding covariate x_2 we are left with a variance for the reduced model of $Var[\hat{\boldsymbol{y}}^R] = \beta_1^2$. As long as $Cov(x_1, x_2) \neq -\frac{\beta_2}{2\beta_1}$, the outcome of the reduced model has an altered variance compared to the complete model. Due to the altered bias and variance, a dichotomizing threshold t that was applied to the complete model cannot be used in the reduced model. Based on the assumption that both the complete and the reduced model should result in an unaltered optimal proportion of high-risk patients, the dichotomizing threshold can be reset such that the proportion of patients classified as high-risk in the training set is equal for the reduced and complete models. However, depending on the information lost by discarding covariates, this assumption of equal proportion may be invalid. In conclusion, the reduced model is likely to be biased, has an altered variance and a reduced accuracy, resulting in an erroneous interpretation.

2 Reweighted model

When applying a risk model with discarded covariates, the discarded covariates may not be correlated to other variables in the model. However, often there is a non-zero covariance between covariates in a model allowing the contribution of discarded covariates to be modelled by the remaining covariates. I.e. giving new weights to the remaining covariates, mimicking the complete model.

If we assume that the training set data is known, then missing covariates can be expressed - at least in part - as a function of the known covariates. We assume there are more training subjects than non-missing covariates to avoid an ill-defined covariance structure.

Let $\boldsymbol{X}$ be the $n \times u$ design matrix containing n patients, a single column for the unit intercept vector and the $u - 1$ non-missing covariates. Let $\boldsymbol{Z}$ be the $n \times v$ response matrix for the same n patients and v missing covariates. Each covariate $z_{.j}$ can be expressed as a function of $\boldsymbol{X}$ by fitting the linear least square regression $\boldsymbol{\Omega} = \left(\boldsymbol{X}^T\boldsymbol{X}\right)^{-1}\boldsymbol{X}^T\boldsymbol{Z}$. This gives the $v \times u$ weight matrix $\boldsymbol{\Omega}$ in which column j contains the weights to be applied

to $\boldsymbol{X}$ in order to obtain the unbiased least square fit onto $z_{.j}$. These weights can be used to redistribute the weights of the missing covariates $\beta(v)$ over the non-missing covariates $\beta(u)$ into $\phi = \beta(u) + \boldsymbol{\Omega}\beta(v)$. Note that in this notation, there is an intercept in $\beta(u)$ and ϕ. Now we determine the reweighted model outcome as $\boldsymbol{y}^R = \boldsymbol{X}\boldsymbol{\phi} + \epsilon$ in which some error will arise due to sampling of the training set.

3 Proof of concept

To show a proof of concept for the reweighted model, we define a complete model $\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta}$ and a reweighted model $\boldsymbol{y}^R = \boldsymbol{X}^R\boldsymbol{\phi}$. The weights $\boldsymbol{\beta}$ are a vector of length m containing randomly drawn values from $\mathcal{N}(\mu = 0, \sigma = 1)$ and $\boldsymbol{X}$ is a multivariate normal distributed $n \times m$ matrix with values drawn from $\mathcal{N}(\mu = 0, \Sigma)$. To construct Σ we initially set the diagonal elements to 1 and randomly draw the off diagonal element from $\mathcal{U}(-\rho, \rho)$. The nearest positive definite matrix is created by the *nearPD* function from the **R Matrix** package.

We set $m = 92$, $n = 350$ and generate two independent matrices $\boldsymbol{X}_{train}$ and $\boldsymbol{X}_{test}$ with the same covariance structure. Next, 36 covariates are randomly selected to be assigned as missing. Based on the training set we determine the vector of reweighted weightings ϕ which are validated by their application to the independently generated matrix $\boldsymbol{X}_{test}$.

In Figure 1 the complete, reduced and reweighted model outcomes are plotted for increasing correlations. A sensitivity analysis of these results is performed by repeating the above procedure 1000×. The correlations and intra-class correlations between the incomplete and complete models obtained are shown in Figure 2. The outcomes of the complete model correlate better to outcomes of the reweighted model than to those of the reduced model, as long as the missing covariates have shared variation to the non-missing covariates.

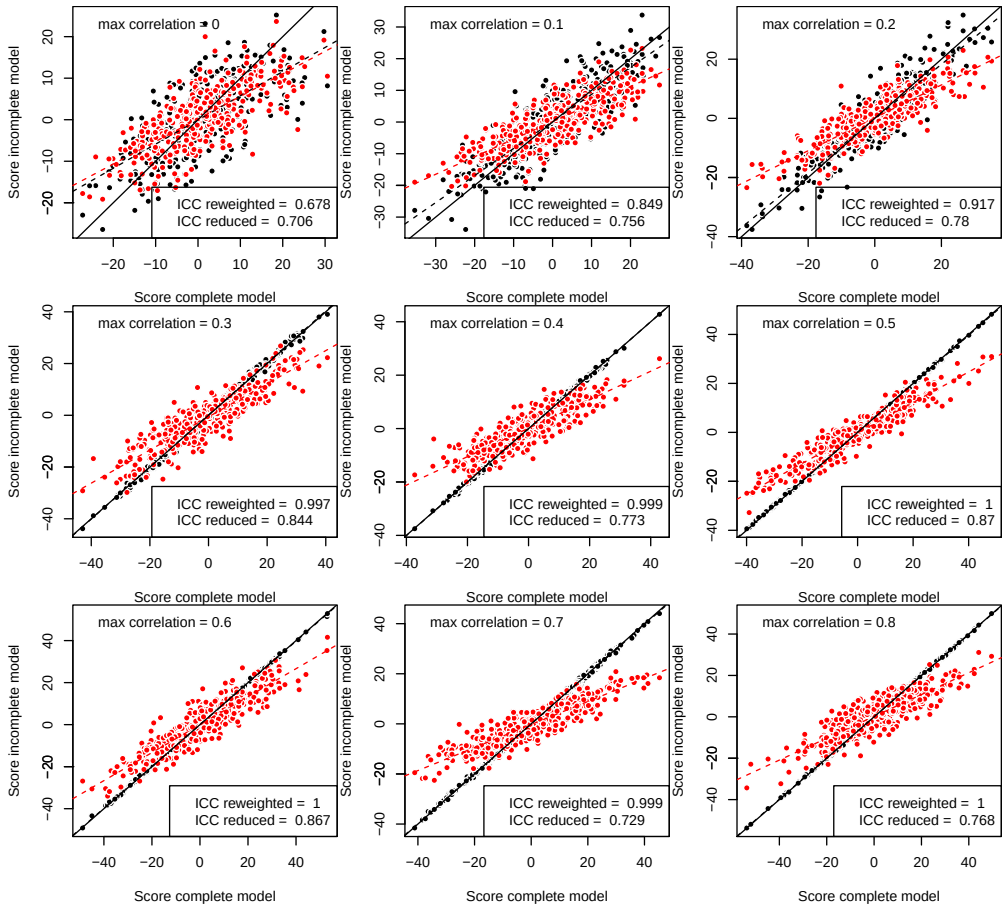

Figure 1: Scatter plots of model outcomes in 9 randomly generated datasets with 350 subjects for ρ (which determines the strength of correlation between covariates) ranging from 0 to 0.8. On the horizontal axes the scores for the complete models with 92 probe-sets are given. On the vertical axes the scores for the reduced (in red) and reweighted (in black) models with 36 missing probe-sets are given.

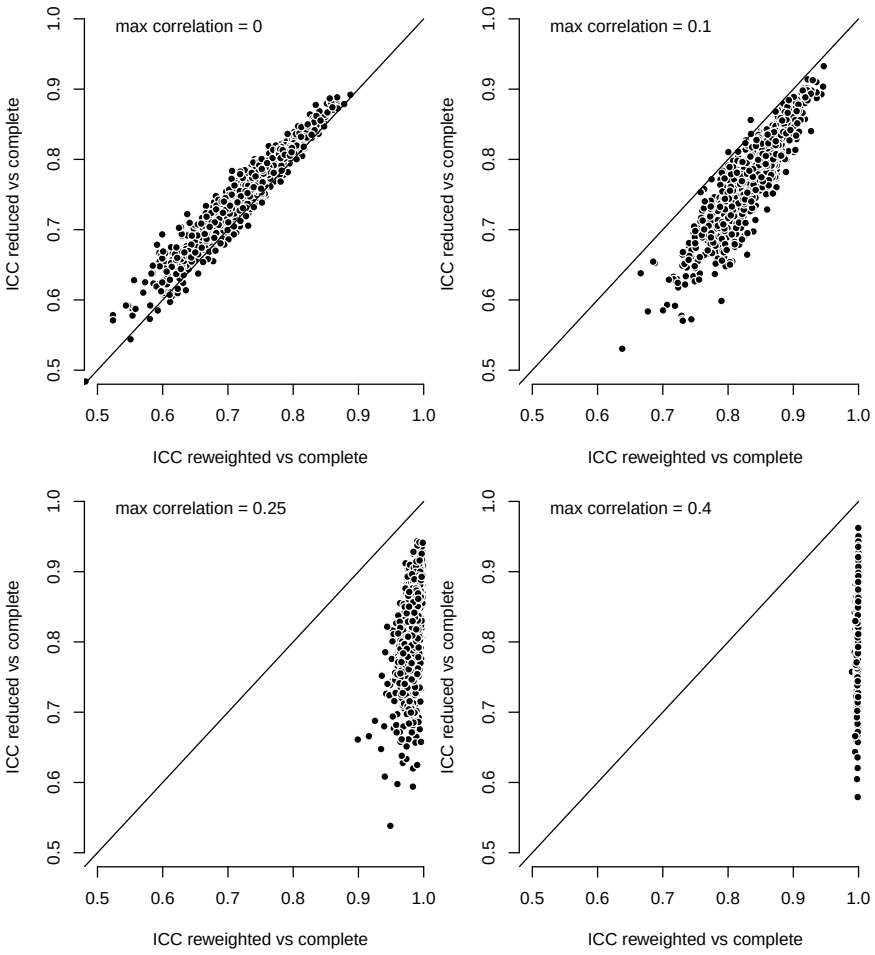

Figure 2: Scatter plots of intra-class correlations (ICC) between the complete and reweighted models (x-axis) and the complete and reduced models (y-axis) as applied on the test data for 1000 independently generated datasets and model weights for $\rho = 0, 0.1, 0.25$ and 0.4. The complete models are based on 92 probe-sets while the incomplete models are based on the same data-set but with 36 randomly assigned missing probe-sets.

4 EMC92

To compare the effect of recalculating weights versus excluding covariates completely in a more realistic setting, we make use of the HOVON65/GMMGHD4 training set and the UAMS-TT2 and UAMS-TT3 test sets which were both profiled using the Affymetrix HG U133 PLUS 2.0 microarray. Expression values within these sets were log2 transformed, mean centered and scaled to unit variance (per probe-set).

4.1 Simulation

- Step1: Randomly select 36 probe-sets out of the EMC92 which are to be missing in the test sets.
- Step2: Determine ϕ for this situation based on the training set.
- Step3: Apply the complete, the reweighted and the reduced model to both test sets.
- Step4: Determine the correlation and intra class correlation (ICC) between both the reduced and reweighted versus the complete model outcomes.

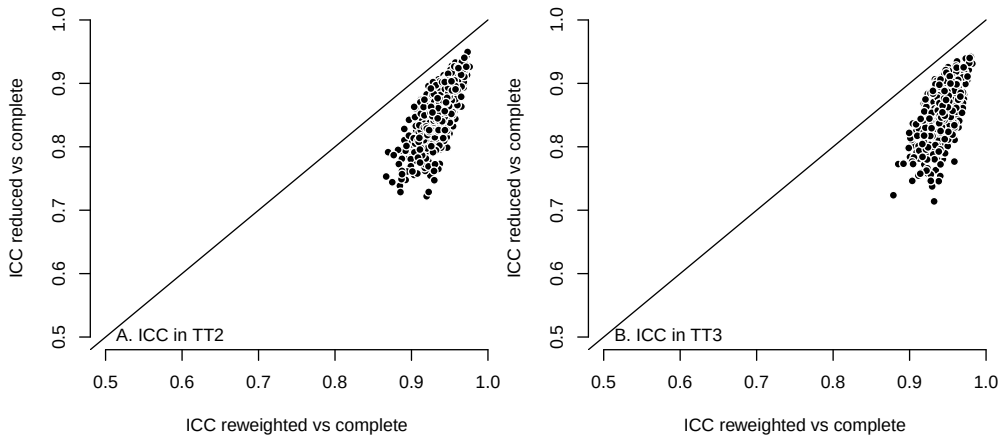

Figure 3: Scatter plots of intra-class correlations (ICC) between the complete and reweighted models (x-axis) and the complete and reduced models (y-axis) as applied on the UAMS-TT2 (left) and UAMS-TT3 (right) for 1000 randomly selected subsets of 36 assumed to be missing probe-sets out the 92 probe-sets of the EMC92-gene classifier.

5 Conclusion

The reweighted model demonstrated a better correlation to the complete model compared to the reduced model. Only in the exceptional case of no correlation among the covariates, did we observe no advantage of the reweighted model (Figure 1). This was observed in the simulation as well in the real data.