ORIGINAL RESEARCH ARTICLE

Prediction of tumour pathological subtype from genomic profile using sparse logistic regression with random effects

Özlem Kaymaz[a], Khaled Alqahtani[b], Henry M. Wood[c] and Arief Gusnanto[d]

[a]Department of Statistics, University of Ankara, Ankara, Turkey; [b]Department of Mathematics, College of Science and Humanitarian Studies, Prince Sattam Bin Abdulaziz University, Al Kharj, Saudi Arabia; [c]Leeds Institute of Medical Research at St. James's, University of Leeds, Leeds LS9 7TF, UK; [d]Department of Statistics, University of Leeds, Leeds LS2 9JT, UK.

ARTICLE HISTORY
Compiled February 3, 2020

ABSTRACT
The purpose of this study is to highlight the application of sparse logistic regression models in dealing with prediction of tumour pathological subtypes based on lung cancer patients' genomic information. We consider sparse logistic regression models to deal with the high dimensionality and correlation between genomic regions. In a hierarchical likelihood (HL) method, it is assumed that the random effects follow a normal distribution and its variance is assumed to follow a gamma distribution. This formulation considers ridge and lasso penalties as special cases. We extend the HL penalty to include a ridge penalty (called 'HLnet') in a similar principle of the elastic net penalty, which is constructed from lasso penalty. The results indicate that the HL penalty creates more sparse estimates than lasso penalty with comparable prediction performance, while HLnet and elastic net penalties have the best prediction performance in real data. We illustrate the methods in a lung cancer study.

KEYWORDS
Tumour; Lung cancer; Pathological subtype; Logistic regression; Sparse solution; Hierarchical likelihood;

1. Introduction

Copy number alterations (CNA) are structural variation in the human genome, in which some regions exhibit more ('gains') or less ('losses') number of copies than the normal two copies. These structural variations are common to be found in cancer patients. Our main interest in this study is to investigate statistical models to predict two different subtypes in lung cancer: squamous carcinoma (SC) and adenocarcinoma (AC)[20]. In the context of cancer patients' care and well-being management, identification of the correct pathological subtype is critical as the treatment administered to them depends on the subtype [2, 4]. Each tumour subtype has different patterns of CNA because the underlying process of their development is different due to the origin of cancer [13]. Therefore, the information contained in CNA is important for prediction of the tumour subtypes.

CONTACT A. Gusnanto. Email: a.gusnanto@leeds.ac.uk

CNA can be estimated using different technologies, such as array competitive genomic hybridization (aCGH) and next-generation sequencing (NGS), the latter of which was used in our motivating dataset (Section 2.1). A CNA dataset typically has some characteristics that can be considered as a challenge from statistical modelling view point. The first one is that the number of genomic regions (variables) far exceeds the number of patients (observations). Secondly, the data exhibit 'blocks' of correlation between genomic regions, because CNA's tend to occur in segments. With such a large number of genomic regions, not all of them are informative in the discriminating the two tumour subtypes. For example, in leukemia, Forero-Castro *et al.* [7] identified that some regions in chromosomes 1, 14, and 15 are important in the identification of cancer subtypes. Brennan *et al.* [3] identified some regions in chromosomes 4, 7, and 12 to be informative to distinguish subtypes in glioblastoma (a type of brain cancer). It is therefore of our interest to develop a model to predict tumour pathological subtypes using the CNA data, while at the same time identify important regions that discriminate the two subtypes.

For this purpose, we consider logistic regression model with a sparse solution. The term 'sparse' refers to the case where some of the model parameters are zero estimated, while the other parameters are estimated to be away from zero. In effect, a variable selection is embedded in the model, since the contribution of some variables for prediction is negated. Tibshirani[22] proposed a penalised regression model with an L1 penalty called lasso. Zou and Hastie[25] identified that the lasso method can have some drawbacks, especially when the data are correlated. Lasso model tends to pick up just one variable (to have non-zero estimate) among the correlated variables. To deal with this challenge, Zou and Hastie[25] introduced the elastic net penalty ('Enet penalty'), in which the lasso penalty is extended by incorporating ridge penalty[10].

Lee and Oh[17] recently proposed a general formulation of random effects model. In a hierarchical likelihood (HL) formulation, the model parameters are assumed to follow a normal distribution, and its variance is assumed to follow a gamma distribution. This formulation creates a penalty function ('HL penalty') that can produce a sparse solution. Furthermore, since the lasso penalty is a special case in this formulation, we consider the HL penalty as an alternative to the lasso penalty. We further extend the HL penalty to incorporate a ridge penalty, similar to the extension of lasso penalty to the elastic net penalty, and call it HLnet (hierarchical likelihood-net) penalty. We apply these methods in a real CNA dataset on lung cancer patients and simulated datasets. The results indicate that the methods produce sparse solutions, with HL penalty create a more sparse solution than lasso penalty. A cross validation on real dataset also indicates that the extension of ridge penalty on the HL penalty improves the model performance by lowering classification error.

We outline the paper as follows. Section 2 describes the methodology involved; Section 3 presents the results and Section 4 contains discussions and concluding remarks.

2. Methodology

2.1. Dataset

The dataset in our study comes from 76 lung cancer patients from Leeds Teaching Hospital (UK), comprising of two groups: squamous carcinoma (38 patients) and adenocarcinoma (38 patients). Details on biological sample preparation, DNA extraction and sequence preparation are described by Wood *et al.*[23]. DNA samples are sequenced using Illumina GAIIx sequencer to produce short DNA sequences. These short sequences, usually called ‘reads’, were aligned using the software package bwa suite version 0.5.9-r16 [18] against assembly hg19 of the human reference genome. We only consider reads that could be uniquely mapped to the human reference genome with quality score ≥ 37.

The copy number alteration (CNA) data from each lung tumour is calculated by ‘depth of coverage’ from their sequences, which basically counts the number of short DNA sequences that are mapped to a fixed-size genomic region (‘window’). For this purpose, the optimal window size for this group of samples is estimated using *NGSoptwin* package to be 150 kbp [8]. The sequence data from 76 cancer patients are not directly comparable because inevitably the tumour samples are contaminated with normal cells by different degrees. To deal with this problem, we performed a normalisation using the *CNAnorm* package [9] to obtain CNA estimates. An example of CNA estimates is presented in Figure 1 for patient LS93, where CNA is estimated as smooth segmented lines [11].

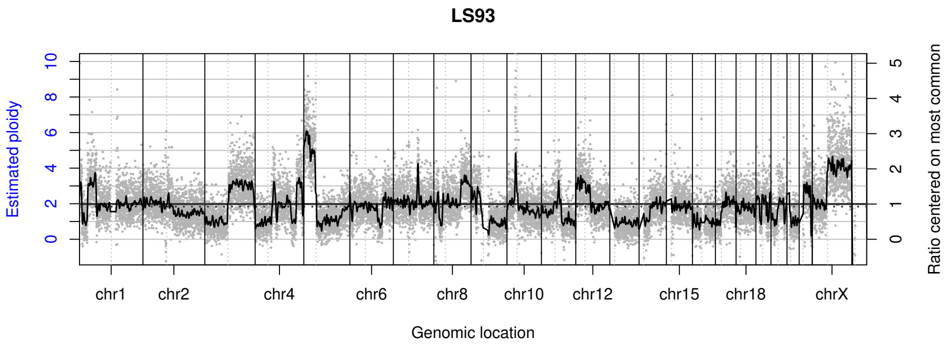

Figure 1. *CNA estimates from one patient (LS93) as smoothed segmented line (solid line) from normalised ratio (grey points). Vertical solid lines separate different chromosomes. The horizontal axis corresponds to the genomic location. I.e. the CNA estimates are ordered from the first to the last position in each chromosome.*

In the dataset that we analyse, we have CNA estimates from 17,571 genomic regions. The CNA estimates from the patients (‘profiles’) are then summarised in a matrix of size 76 by 17,571 after exclusion of sex chromosomes and the centromere regions that contain many missing values. The columns of the matrix are ordered based on the genomic locations. One of the characteristics of the data is the presence of correlation ‘blocks’ as shown in Figure 2. The figure shows the correlation in chromosomes 1 and chromosome 2.

2.2. Sparse logistic regression

Let y be a vector of the tumour subtype, where $y_i = 1$ if the i-th tumour is of squamous carcinoma (SC) subtype and $y_i = 0$ if it is of adeno carcinoma (AC), for $i = 1, 2, \ldots, n$. Let X be a $n \times p$ matrix of fixed predictors that contains the column for intercept. In general, any important clinical predictors can be included in X. However, in our application, it is currently associated with a fixed intercept, i.e. X is a vector of ones. Let β be a p-vector of fixed parameters associated with X. The CNA data matrix is denoted as Z of size $n \times q$ with a q-vector of random parameters b, $b \equiv \{b_j\}$, $j = 1, 2, \ldots, q$. We assume that each y_i follows a Binomial distribution with mean μ_i,

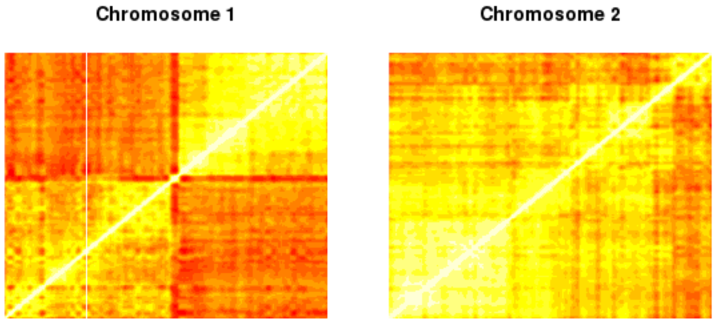

Figure 2. *Correlations between genomic regions in the CNA dataset across 76 lung cancer patients, in chromosomes 1 (1,469 windows) and chromosome 2 (1,561 windows). The centromere regions that contain missing values have been excluded. In the whole genome (excluding sex chromosomes), the correlations range from -0.47 to 0.999, with mean(median) of absolute correlation to be 0.28(0.27). Bright colours indicate high positive correlation.*

or $y_i \sim \text{Bin}(1, \mu_i)$. We model the mean μ_i, in vector notation, as

$$h(\mu) = X\beta + Zb, \tag{1}$$

where $h(\cdot)$ is a logit link function that applies element wise, $\mu \equiv \{\mu_i\}, i = 1, 2, \ldots, n$, and we assume that b follows a distribution, which will be described below. The term $X\beta$ in the above equation can be replaced by β_0 in our current specific application because it is just a fixed intercept. However, we keep the above formulation in the equation for other general applications.

The formulation of the above model (1) follows that of generalised linear models with random effects [15]. In a general context, the random effects b are usually assumed to follow a normal distribution with mean zero and variance $D(\theta)$, or $b \sim N(0, D(\theta))$, where $D(\theta) \equiv \theta I_q$ for a tuning parameter θ and I_q is the identity matrix of size q. However, it is well know that this assumption on b will produce a ridge estimate, which is not a sparse solution.

To obtain a sparse solution, Tibshirani [22] proposed a lasso penalty, which is equivalent to assuming a Laplace distribution on the random effects. One main drawback of the lasso model is that the when dealing with a block of correlated variables, which is common in our data, lasso will pick one variable to have non zero estimate in that block [25]. To address this problem, Zou and Hastie [25] proposed elastic-net model, which is equivalent to assuming a mixture distribution of normal and Laplace distribution on the random effects b. This assumption produces a sparse solution with some 'grouping effect' on correlated variables [25].

In this study, we consider the random effects proposed by Lee and Oh [17], as an alternative to the lasso and elastic-net models, to produce sparse solution in the above model. Lee and Oh [17] proposed that the random effects b is assumed to follow a normal distribution such that

$$b_j|u_j \sim N(0, u_j\theta) \tag{2}$$

where θ is a tuning parameter with $\theta = 2\sigma^2$, and u_j is assumed to follow gamma distribution with parameter w

$$p_w(u) = \left(\frac{1}{w}\right)^{\frac{1}{w}} \frac{1}{\Gamma\left(\frac{1}{w}\right)} u^{\frac{1}{w}-1} \exp\left(-\frac{u}{w}\right)$$

and $E(u) = 1$ and $\text{Var}(u) = w$.

To use the hierarchical likelihood approach [15, 16], Lee and Oh [17] reformulated the above model (2) by noting that it can be written as

$$b_j = \sqrt{\tau_j} e, \tag{3}$$

where $\tau_j = 2\sigma^2 u_j$ and e follows a standard normal distribution, i.e. $e \sim N(0, 1)$. This means $\log(\tau_j) = \log 2\sigma^2 + v_j$ where $v_j = \log u_j$.

Following [17], the joint log likelihood of the parameters in the logistic regression (1) can then be expressed as hierarchical likelihood [15, 16]

$$\ell(\beta, \theta, b) = \log p(y|b) + \log p(b), \tag{4}$$

where

- $\log p(y|b)$ is the log likelihood of the logistic regression model (1)

  $$\log p(y|b) = \sum_{i=1}^{n} \{y_i \log \mu_i + (1 - y_i) \log(1 - \mu_i)\},$$

  with $\mu_i = \left(1 + \exp\{-(X_i^T \beta + Z_i^T b)\}\right)^{-1}$, and X_i and Z_i are respectively the i-th row of X and Z,
- $\log p(b)$ is the hierarchical log likelihood of the random effects from the definition in Eq. (2) and defined as

  $$\log p(b) = \sum_{j=1}^{q} \log p(b_j|u_j) + \sum_{j=1}^{q} \log p(\log u_j) \tag{5}$$

  where

  $$\log p(b_j|u_j) = -\frac{1}{2} \left\{ \log\left(4\pi\sigma^2\right) + \log u_j + \frac{b_j^2}{2\sigma^2 u_j} \right\}$$

  and, by noting that $v_j = \log u_j$,

  $$\log p(v_j) = \log p(\log u_j) = -\frac{\log(w)}{w} - \log \Gamma\left(\frac{1}{w}\right) + \frac{\log u_j}{w} - \frac{u_j}{w}.$$

2.3. Parameter estimation

Given (w, θ), the estimation of the parameters β and b is done using the profile h-likelihood [15–17]

$$\ell_p(\beta, \theta, b) = \log p(y|b)|_{u=\hat{u}} + \log p(b)|_{u=\hat{u}} \,. \tag{6}$$

Lee and Oh [17] showed that, under the above formulation, u is estimated as a solution to $d\ell/du = 0$ and given by

$$\hat{u}_j = \hat{u}(b_j) = w\left\{(2/w - 1) + \kappa_j\right\}/4, \tag{7}$$

where $\kappa_j = \sqrt{4b_j^2/(w\sigma^2) + (2/w - 1)^2}$.

In the above formulation, w is not estimated and, given b, the relationship between $\hat{u}$ and w in Eq. (7) is fixed. Therefore, there are only two parameters (β, b) and one tuning parameter θ (note that $\sigma^2 = \theta/2$) to estimate. The type of penalty, hence solution obtained, is characterised by the choice of w. For $w = 0$ (in limit terms), we have [17]

$$\log p(b)|_{w=0} \propto 1/(2\theta) \sum_{j=1}^{q} b_j^2 \tag{8}$$

which gives the ridge penalty. For $w = 2$, we have [17]

$$\log p(b)|_{w=2} \propto 1/\sqrt{\theta} \sum_{j=1}^{q} |b_j| \tag{9}$$

which gives the lasso penalty. The penalty function that we refer to in the above formulation (HL penalty) is $\log p(b)$ when $w = 30$ as suggested by Lee and Oh [17]. The shape of the penalty functions are presented in Figure 3. Relative to the lasso penalty (L1, $w = 2$), the HL penalty ($w = 30$) shows a sharp upside-down spike around zero and quickly spread evenly as b is away from zero. This shape of penalty is expected to produce a very sparse solution because the spike around zero that is more prominent than the lasso penalty, while at the same time only allow estimates that are considered 'large enough'.

The parameter estimation is done within iterative weighted least squares (IWLS) [15, 16, 19]. The estimation of β and b is performed at a fixed θ, and then θ is estimated through cross validation or Akaike's information criterion (AIC, described below). We denote $\lambda = 1/\theta$, and $W = \text{diag}(1/\hat{u}_j)$. At fixed θ, we differentiate the profile likelihood $\ell_p(\beta, \theta, b)$ (Eq. (6)) with regard to each of β and b to obtain the mixed model equations [16, 19]

$$\begin{pmatrix} X^T\Sigma^{-1}X & X^T\Sigma^{-1}Z \\ Z^T\Sigma^{-1}X & Z^T\Sigma^{-1}Z + \lambda W \end{pmatrix} \begin{pmatrix} \beta \\ b \end{pmatrix} = \begin{pmatrix} X^T\Sigma^{-1}Y \\ Z^T\Sigma^{-1}Y \end{pmatrix}$$

where Y is a working vector with elements

$$Y_i = X_i^T\beta + Z_i^T b + \frac{y_i - \mu_i}{\mu_i(1 - \mu_i)}$$

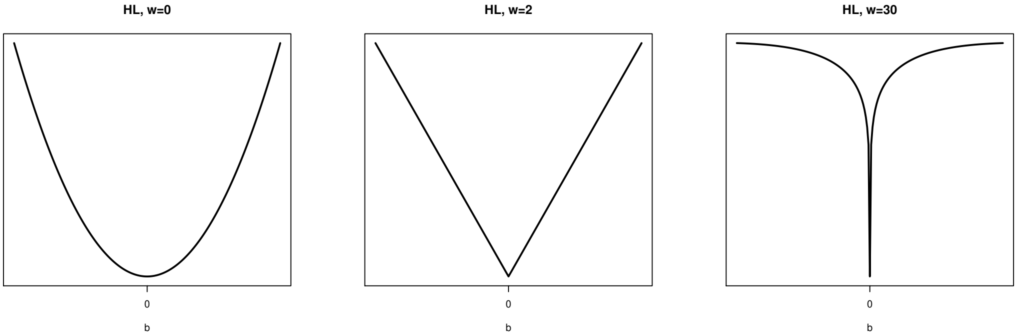

Figure 3. *The shape of penalty function* $\log p(b)$ *under the hierarchical likelihood[17], when* $w = 0$ *(in limit terms) that produces ridge penalty,* $w = 2$ *that produces lasso (L1) penalty, and* $w = 30$ *('HL penalty') for a particular fixed* θ*.*

and Σ is a diagonal matrix with $\Sigma_{ii} = \mu_i(1 - \mu_i)$.

At fixed θ and using starting values of $\widehat{\beta}$ and $\hat{b}$, we estimate β and b by iterating

$$\widehat{\beta} = (X^T\Sigma^{-1}X)^{-1}X^T\Sigma^{-1}(Y - Z\hat{b}) \tag{10}$$

and

$$\hat{b} = (Z^T\Sigma^{-1}Z + \lambda W)^{-1}Z^T\Sigma^{-1}(Y - X\widehat{\beta}) \tag{11}$$

where β, b, and μ are evaluated at the current values. In the above estimation of b, the value of u_j in the diagonal of W can go to zero in the iteration when the corresponding estimate of b_j is close to zero. It is important to note that, when the estimate of b_j is going close to zero in the iteration, the corresponding u_j in the diagonal of W is also going to zero. To avoid failure in the matrix inversion, we add a small constant 10^{-8} to each u_j. This means that the estimate b_j is considered zero when their magnitude is less than, say, 10^{-5} or 10^{-4}.

The estimation of $\lambda \equiv 1/\theta$ can be done by performing a cross-validation (see Section 2.5) on different values of λ and choose the one that minimise classification error. Alternatively, by calculating AIC=$-2\log p(y|b) + 2$df where

$$\text{df} = \text{trace}\left\{(Z^T\Sigma^{-1}Z + \lambda W)^{-1}Z^T\Sigma^{-1}Z\right\}$$

across different value of λ's and select the one that minimises AIC[19].

2.4. Extension

A simple and natural extension to the above model are by incorporating the ridge penalty to the HL penalty. Zou and Hastie[25] incorporate both the lasso penalty (9) and ridge penalty (8) in the model likelihood to create the so called 'elastic net' that (from the formulation of Lee and Oh[17]) can be written as

$$\log p_{\text{enet}}(b) \propto 0.5\,\log p(b)|_{w=2} + 0.5\,\log p(b)|_{w=0}$$

following Eq. (8) and (9). Their proposal was motivated by the drawback of lasso penalty in the context of correlated data: the solution will pick up a variable in a 'block' of correlated variables and it does not care which one. By incorporating the ridge penalty, they expect that the 'grouping effect' [25] of the ridge penalty will improve this situation.

Similarly, we can extend further the HL penalty to incorporate ridge penalty with equal weight in dealing with correlated data. In this case the penalty function becomes

$$\log p_{\text{HLnet}}(b) \propto 0.5 \log p(b)|_{w=30} + 0.5 \log p(b)|_{w=0} \tag{12}$$

where the first term corresponds to the HL penalty and the second one to the ridge penalty (8). For simplicity, we refer the above Eq. (12) as 'HLnet' penalty. As described above, the HL penalty is expected to produce a very sparse solution given the shape of the penalty function. The incorporation of the ridge penalty in the HLnet penalty (12) is expected to benefit from its 'grouping effect' [25] in the solution when dealing with correlated data such as our genomic dataset. This extension can be considered as assuming a type of mixture distribution on random effects b as described further in the Supplementary Material, along with some description on the parameter estimation. In the above formulation, the weight for each penalty is set to be 0.5 by default, although in practice this is specified by user. The motivation to consider the weight of 0.5 is described further in the Supplementary Material.

2.5. Cross validation

We consider cross-validation (CV) as an alternative in estimating λ and as a method to estimate the model performance in terms of classification error in prediction based on 'new samples'. From n observations in the data, we randomly split them into training set of size n_t and validation set of size n_v ($n_t + n_v = n$) such that

$$y := \begin{bmatrix} y_t \\ \cdots \\ y_v \end{bmatrix}, X := \begin{bmatrix} X_t \\ \cdots \\ X_v \end{bmatrix}, Z := \begin{bmatrix} Z_t \\ \cdots \\ Z_v \end{bmatrix}.$$

The training set serves as the set by which we estimate the model parameters $\widehat{\beta}_t$ and $\widehat{b}_t$. The estimates are then used in the validation set to obtain model prediction

$$\widehat{y}_v = I\left(h^{-1}\left\{X_v\widehat{\beta}_t + Z_v\widehat{b}_t\right\} \geq 0.5\right), \tag{13}$$

where $I(\cdot)$ equals one (squamouscarcinoma) if the expression inside the brackets is true, and zero (adenocarcinoma) otherwise.

Given the predicted group labels $\widehat{y}_v$, we calculate the classification error in the validation set by comparing it with the observed group labels y_v. Let us denote $y_v = (y_{v1}\ y_{v2}\ \ldots\ y_{vn_v})^T$ and, from (13) $\widehat{y}_v = (\widehat{y}_{v1}\ \widehat{y}_{v2}\ \ldots\ \widehat{y}_{vn_v})^T$, we define the classification error as

$$CE = \frac{1}{n_v}\sum_{k=1}^{n_v} I(y_{vk} \neq \widehat{y}_{vk}). \tag{14}$$

In practice, we performed a five-fold cross validation and the classification error is calculated in the validation sets across the five folds. This is done for each model that we consider in this study, and the optimal λ (in each model) is estimated as the one that minimises the (five-fold) classification error.

2.6. Simulation study

To understand the working characteristics of the model, we consider a simulation study under the following setting. To mimic the real CNA data, we generate a matrix of CNA data Z of size $n = 100 \times 1000 = q$ as $Z \sim MVN(0, \Psi)$ where MVN is the multivariate normal distribution density and Ψ is 1000×1000 covariance matrix. For a known K, it is defined as

$$\Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} & 0 & \cdots & 0 \\ \Psi_{21} & \Psi_{22} & \Psi_{23} & \cdots & 0 \\ 0 & \Psi_{32} & \Psi_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Psi_{KK} \end{bmatrix}. \tag{15}$$

For $k = 1, 2, \ldots, K$, Ψ_{kk} is of size $K' \times K'$, $K' = (1000/K)$, and is defined as

$$\Psi_{kk} = \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & \vdots & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}$$

with $\rho = 0.9$. For $k = 1, 2, \ldots, K-1$, we define $\Psi_{k(k+1)} = \Psi_{(k+1)k}$ of size $K' \times K'$ as

$$\Psi_{k(k+1)} = \Psi_{(k+1)k} = \begin{bmatrix} r & r & \cdots & r \\ r & r & \cdots & r \\ \vdots & \vdots & \ddots & \vdots \\ r & r & \cdots & r \end{bmatrix}.$$

In this simulation study, we consider two different values of r. The first one is $r = 0$, which means that the covariance matrix in Equation (15) is a block diagonal matrix with independent blocks. The second one is $r = 0.4$, which means that the blocks are moderately correlated. The latter setting, i.e. correlated diagonal blocks, is closer to the characteristics that we see in the real data (Figure 2).

In the simulation setting, K is set to be $200, 100$, and 50, corresponding to have a correlation block Ψ_{kk} of size K'=5, 10, and 20 respectively. To get the simulated subtype, we set the true β to be zero (although we still allow an intercept to be fitted in the estimation), and the true random effects b to be set in two different scenarios:

- ***Scenario A*** b_j is set equal to 1 for the first K' variables, -1 for the second K' variables, and zero everywhere else, or

$$b = (\underbrace{1, 1, \ldots, 1}_{\text{First } K' \text{ variables}}, \underbrace{-1, -1, \ldots, -1}_{\text{Second } K' \text{ variables}}, 0, 0 \ldots, 0)^T.$$

In this setting, the truly associated variables corresponds to the first two segments of correlated variables in the simulated data.

- ***Scenario B*** b_j is set equal to 1 for the first variable within the first K' variables, -1 for the first variable within the second K' variables, and zero everywhere else, or

$$b = (\underbrace{1, 0, \ldots, 0}_{\text{First } K' \text{ variables}}, \underbrace{-1, 0, \ldots, 0}_{\text{Second } K' \text{ variables}}, 0, 0 \ldots, 0)^T.$$

In this setting, the truly associated variables are only a 'singleton' in each of the first two segments of correlated variables in the simulated data.

To summarise, we fit five different models in this simulation study: lasso (L1), HL, ridge, elastic net (enet), and HLnet. The ridge model is included in this simulation as a benchmark of non-sparse model. Each model is fitted to simulated data under each of combination of settings: (1) the size of the segments (K') is 5, 10, and 20, (2) the configuration of truly associated variables (scenarios A and B above), and (3) the status of the diagonal blocks in the covariance matrix (whether they are independent or correlated). Hence 12 dataset settings for each model. The models are fitted on 100 simulated datasets for each simulated data setting using their respective tuning parameter estimates (using AIC), and from each of them we calculate (1) classification error from cross validation, and (2) sensitivity and specificity of sparse solution. In the latter, we are interested in the proportion of truly non-zero parameters that are estimated as non-zero (sensitivity) and in the proportion of truly zero parameters that are estimated as zero (specificity).

The high correlation structures that we impose on the simulated data make the simulation very challenging. In such challenging circumstance, the two different scenarios (A and B) represents two different extreme situations: scenario A indicates that the true effects cover the whole correlation block while scenario B indicates that the true effect is only on one variable within a correlation block. Combined with high correlation within the data, identifying the one true effect within the correlation block will be extremely difficult since the other variables within the block are almost equally good.

3. Results

3.1. Simulated data

In this section, we present the results of the simulation study. Figure 4 shows the classification errors for simulated data across different models and simulation scenarios, when the diagonal blocks of the data covariance matrix are independent. Similarly, Figure 5 shows the corresponding classification errors when the diagonal blocks of the data covariance matrix are moderately correlated.

Under scenario A in Figure 4, the classification errors of logistic regression models with lasso penalty (L1) are significantly lower than those with HL and HLnet penalties, across different block sizes (K'). The classification errors of those with HL penalty are not significanty different that those with HLnet penalty. Under scenario B, we find that there are no significant differences of classification errors between those models with L1, HL, and HLnet penalties, except when the block size is 5. In this case, the errors for the HLnet model is significantly lower than those of L1 and HL penalties.

When we consider logistic regression model with elastic net penalty (Enet) in Figure

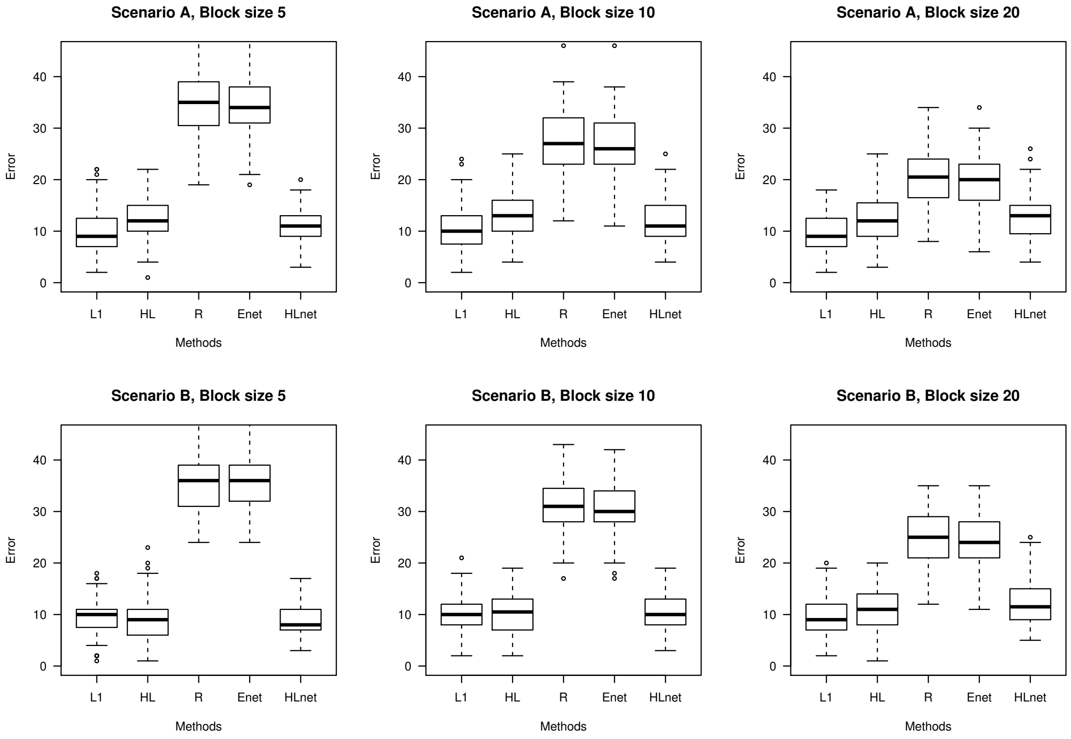

Figure 4. *Classification error (in percent) of logistic regression models with lasso penalty (L1), HL penalty (HL), ridge penalty (R), elastic net penalty (Enet), and HLnet penalty (HLnet) from 100 simulated data under different simulation settings when the diagonal blocks of the data covariance matrix are independent. Details of the simulation settings are described in the main text Section 2.6.*

4, the classification error is as high as that with ridge penalty. Although the sparsity is imposed in the Enet model with the lasso penalty, the figure suggests that the performance of the model is comparable to that with ridge penalty. Only when the HLnet penalty is considered (ridge and HL penalties), the model performance improves substantially and becomes closer to the performance of L1 and HL models. The same conclusion can be drawn for Figure 5.

In Figure 5 where the diagonal blocks of the data covariance matrix are moderately correlated, the classification of errors of the L1 models are also significantly lower than those of HL and HLnet models in scenario A across different block sizes. The errors for HL model and HLnet model, on the other hand, are comparable. However, in scenario B, a different conlusion can be drawn. The errors for HL models are significantly lower than those of L1 models across different block sizes. Relative to errors from L1 models, those from the HLnet models are lower when the bock size is 5, equal when the block size is 10, and higher when the block size is 20.

The sensitivity and specificity of the random effects estimates are presented in Table 1 for the different models in the simulation. Sensitivity refers to the proportion of truly non-zero random effects that are estimated to be non zero, while specificity refers to the proportion of truly zero random effects that are zero estimated. Overall, the sensitivities and specificities of the models are similar between different correlation structure in the data. Within scenario A, in which the true effects cover the whole correlation block of the data, the sensitivity of L1 and HL models tend to decrease as the block size increases. However, the sensitivity for the Enet and HLnet models

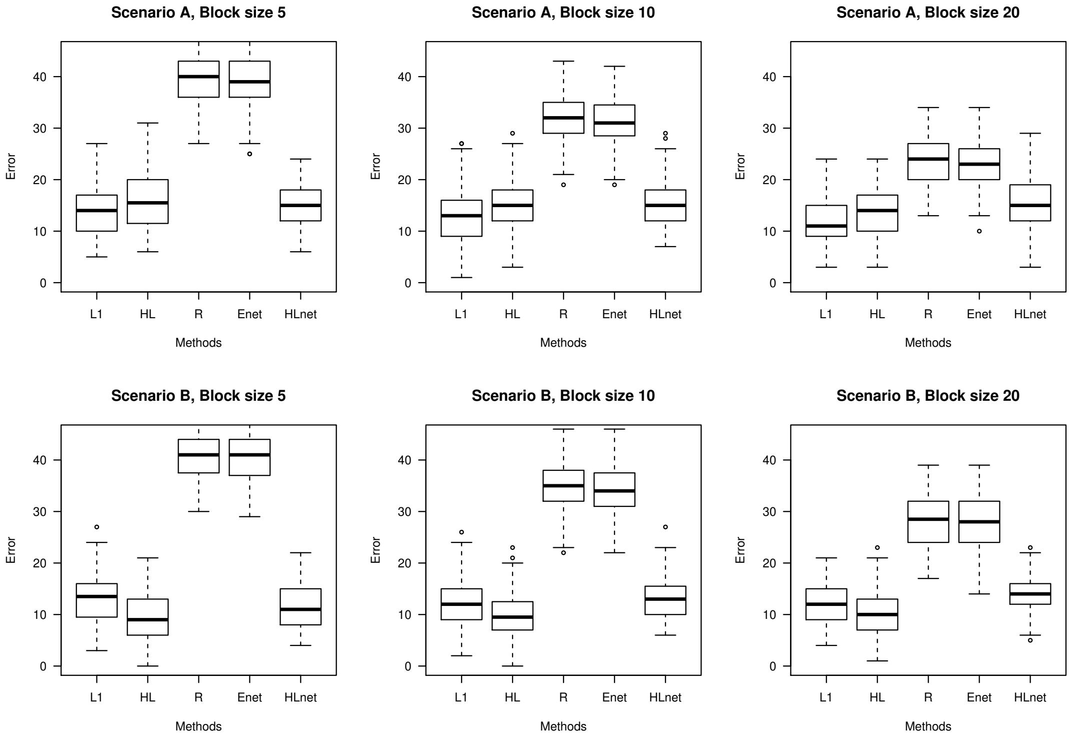

Figure 5. *Classification error (in percent) of logistic regression models with lasso penalty (L1), HL penalty (HL), ridge penalty (R), elastic net penalty (Enet), and HLnet penalty (HLnet) from 100 simulated data under different simulation settings when the diagonal blocks of the data covariance matrix are moderately correlated. Details of the simulation settings are described in the main text Section 2.6.*

Block Size	Methods (penalty)	Independent block				Correlated block			
		Scenario A		Scenario B		Scenario A		Scenario B	
		Sens.	Spec.	Sens.	Spec.	Sens.	Spec.	Sens.	Spec.
5	L1	0.930	0.918	1.000	0.908	0.929	0.910	1.000	0.901
	HL	0.354	0.996	0.980	0.997	0.398	0.995	0.995	0.996
	Enet	0.932	0.875	0.935	0.907	0.612	0.900	0.520	0.912
	HLnet	0.998	0.883	1.000	0.873	1.000	0.799	1.000	0.825
10	L1	0.783	0.930	1.000	0.912	0.774	0.922	1.000	0.908
	HL	0.186	0.996	0.940	0.996	0.219	0.995	0.985	0.996
	Enet	1.000	0.413	1.000	0.363	1.000	0.324	1.000	0.277
	HLnet	0.994	0.816	1.000	0.803	0.999	0.702	1.000	0.669
20	L1	0.563	0.944	1.000	0.925	0.572	0.941	1.000	0.922
	HL	0.099	0.996	0.865	0.996	0.111	0.996	0.985	0.997
	Enet	1.000	0.353	1.000	0.272	1.000	0.308	1.000	0.236
	HLnet	0.999	0.707	1.000	0.610	0.998	0.603	1.000	0.491

Table 1. *Sensitivity and specificity of the different logistic regression models with lasso penalty (L1), HL penalty (HL), elastic net penalty (Enet), and HLnet penalty (HLnet) from 100 simulated data across different simulation settings. The settings are described in the main text Section 2.6. The results of model with ridge penalty are not included as it does not produce sparse solution (i.e. sensitivity of one and specificity of zero).*

tend to be stable and high. Within scenario B, all of the models tend to have a stable and high sensitivity as the block size increases. The sensitivity of the model with lasso penalty (L1) is substantially higher than that of HL penalty. On the other side, the specificity of model with L1 penalty, albeit comparable, is slightly lower than that with HL penalty. Therefore, we can conclude that the model with HL penalty produces a more sparse solution than that with L1 penalty.

While the sensitivity of L1 model remains relatively high when it is extended to Enet penalty, the sensitivity of HLnet model increases substantially in scenario A. Unfortunately, this is at the cost of specificity, as expected. The incorporation of ridge penalty in the Enet and HLnet penalty increases the number of non-zero estimates. Given the high correlation structure in the simulated data, it is very easy for the other variables within a block to have non-zero estimates when one of the variable has non-zero estimate. It is also interesting to note that, as the correlation block size increases, the sensitivity of L1 and HL models decreases substantially compared to those of Enet and HLnet models, while the specificity approximately remains the same. For scenario B, where the true effect is only in one variable within a block of correlated variables, the sensitivity of the sparse methods (L1, HL, enet, and HLnet penalties) are consistently high for different block sizes. The specificity of those methods is also similar to that in Scenario A.

3.2. Lung cancer data

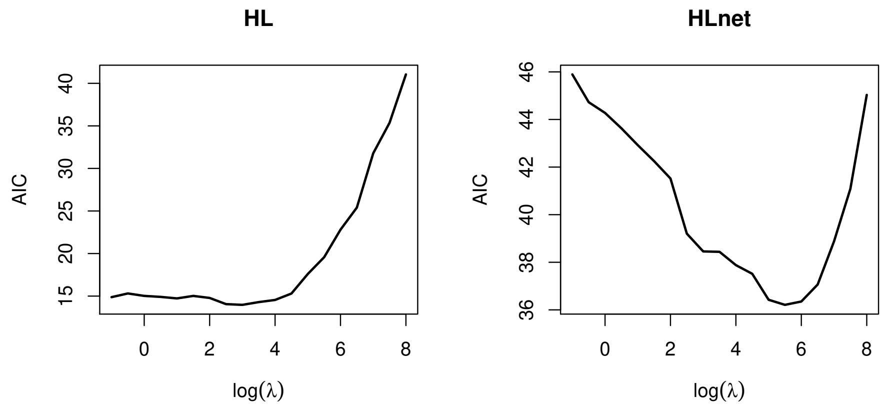

Figure 6. *Illustration of the AIC across different* $\log(\lambda)$ *in the logistic regression models with HL and HLnet penalties.*

We now discuss the results of fitting the models to the lung cancer data. The estimation of λ via AIC is illustrated in Figure 6, with HL and HLnet penalties as example. The figure shows AIC at different values of λ's evaluated and the optimal λ's for the HL and HLnet models are obtained at exp(3) (≈ 20.1) and exp(5.5) (≈ 244.7), respectively.

The random effects estimates $\widehat{b}$ for the different models are presented in Figure 7, at their respective optimal λ's. The figure indicates that the logistic regression models with lasso (L1) and HL penalties produce sparse solutions. The model based on L1 penalty produces 319 non-zero estimates (approximately 1.8% out of 17,571 genomic regions), while those based on the HL penalty produces only 10 non-zero estimates. Seven non-zero estimates are common for the two methods. This indicates that the HL penalty produces very sparse solution, relative to that of L1 penalty. The estimates of the model with lasso penalty creates 29 groups. A group here is defined as consecutive variables with non-zero estimates. As the variables correspond to physical locations in the genome, then the groups correspond to consecutive regions in the genome. On the other hand, being very sparse, the estimates from HL penalty come from single regions.

Figure 7 also shows the estimates of the model with elastic net (Enet) and HLnet penalties. The estimates from the model with HLnet penalty are still sparse, with eight non-zero estimates. In terms of number of non-zero estimates, this is less than that from HL penalty. However, these non-zero estimates (under HLnet penalty) are associated with four groups (consecutive variables) of genomic regions. On the other hand, the Enet penalty has produced 1,231 zero-estimates (approx. 7%) and the other 93% estimates are non-zero. The expected sparse solution of the estimates is not fully achieve, indicating that the solution is closer to the ridge penalty (at optimal λ based on AIC).

To investigate the predictive ability of the models, we consider five-fold cross validation and calculate the classification error at different values of λ. The results of the cross validation are presented in Figures 8. Overall, the models with elastic net and HLnet penalties have the lowest classification error. One thing to note is that the models with L1 and HL penalties has classification error that is very dependent on the estimate of λ, which those with Enet and HLnet penalties have relatively low classification error at a broad range of λ.

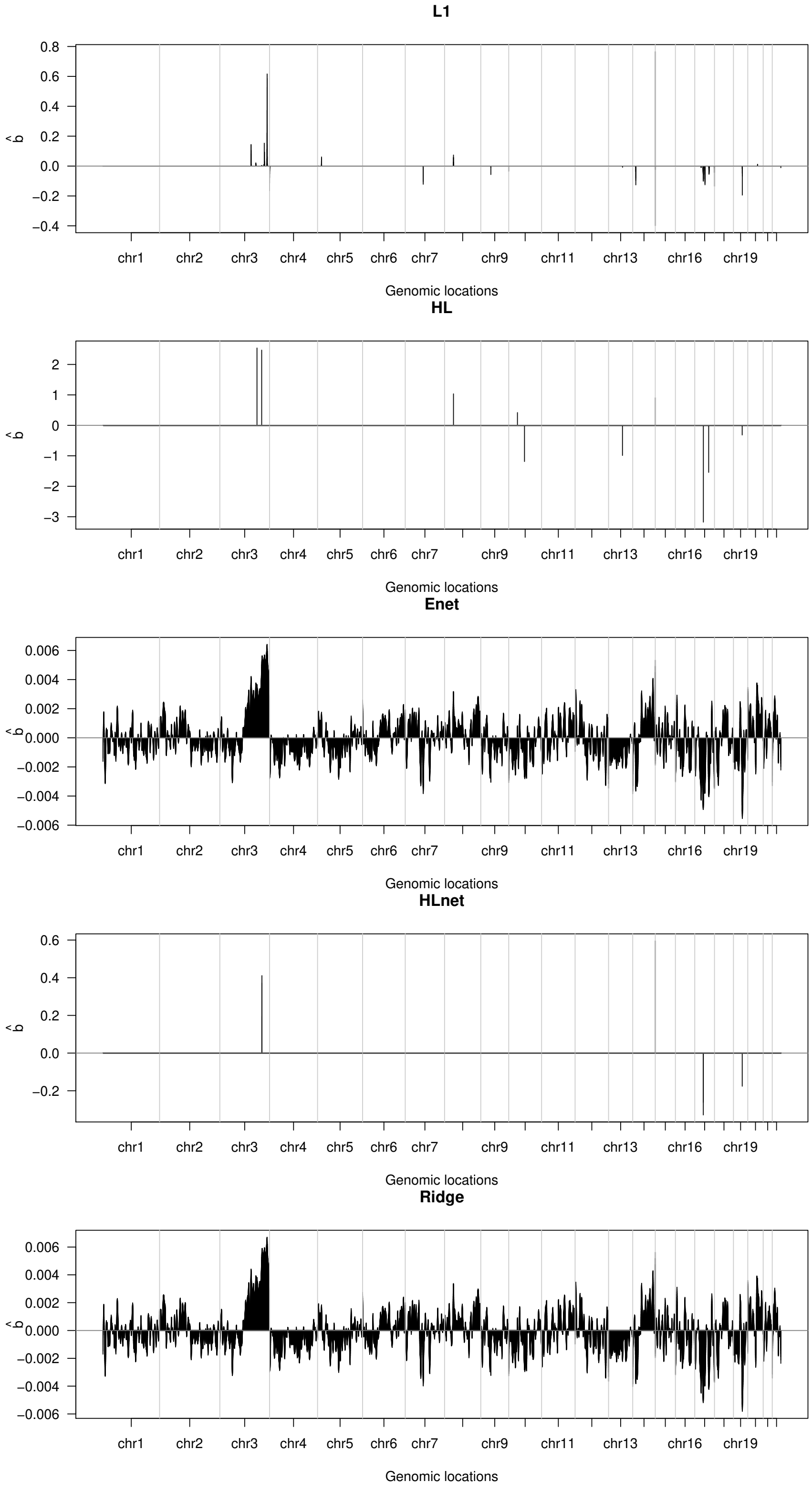

Figure 7. *Estimates of the random effects, $\widehat{b}$, in the logistic regression with lasso (L1), HL, elastic net (Enet), HLnet, and ridge penalties (at their respective optimal λ based on AIC).*

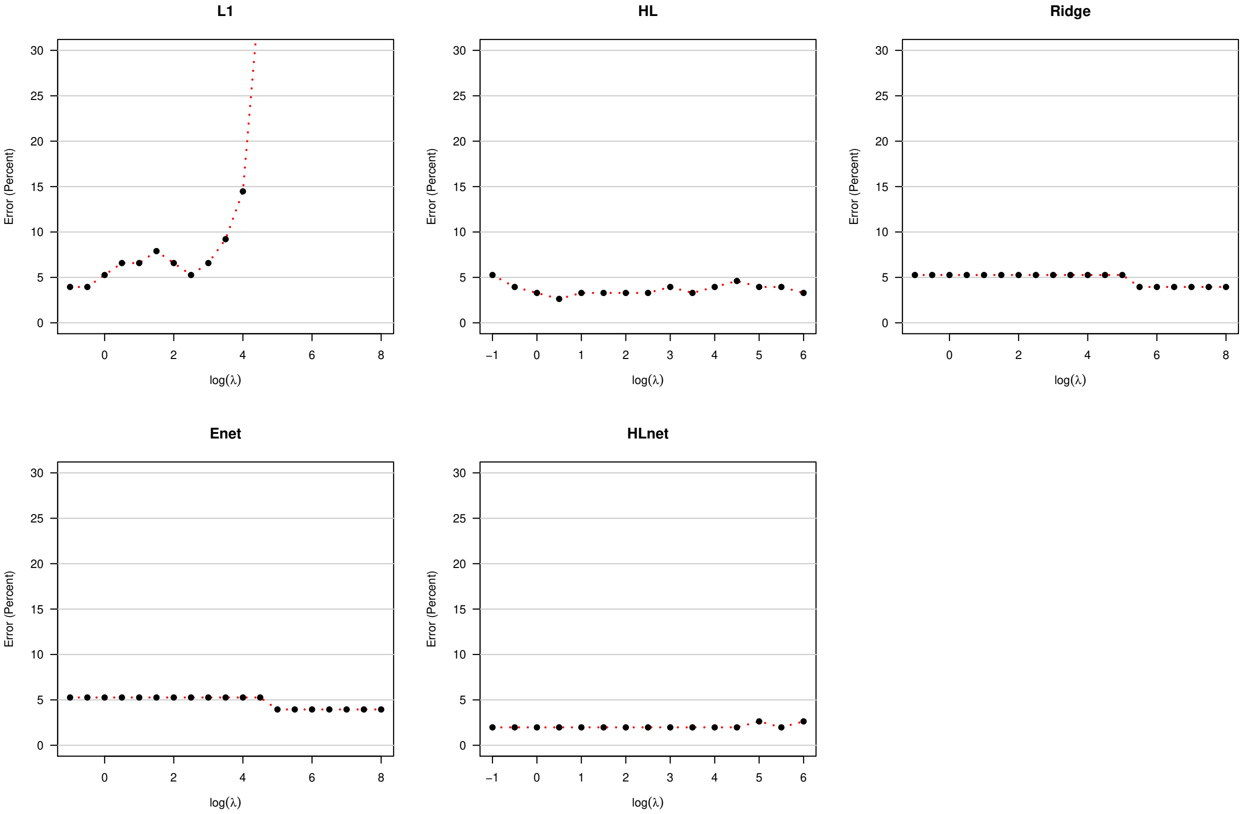

Figure 8. *Classification errors (in percent) of logistic regressions with lasso (L1), HL, elastic net (Enet), HLnet, and ridge penalties in a five-fold cross validation of the lung cancer dataset.*

4. Discussion and concluding remarks

In the context of personalised medicine of lung cancer patients, a main objective in the statistical modelling is accurate prediction of tumours' subtypes[14]. This type of modelling often involves analysis of high-dimensional data and sparse solution is desired in general due to its variable selection effect that makes it easier to identify important variables responsible for prediction[1, 5]. We have considered the application of HL penalty as an alternative to the lasso (L1) penalty for prediction based on genomic data as both models are specific cases of the model formulation by Lee and Oh[17].

We recognise that this formulation has some advantages. The first one is that the estimation of the parameters can be done using iterative weighted least squares (IWLS). This is relatively easy to implement compared to, for example, the lars algorithm[6], gradient descent[24], or Gibbs sampling[21]. Secondly, incorporation of other suitable penalty function can be taken into account in IWLS. An immediate extension is the incorporation of ridge penalty as described above to create HLnet penalty. Another example is the incorporation of penalty function from Cauchy distribution that, with a smoother matrix, creates a smooth estimate of b[12]. Further extension to incorporate dependencies between genomic regions is currently our active research and beyond the scope of this manuscript.

Analysis on the real and simulated data indicate that the HL penalty creates a more sparse solution than L1 penalty. Our simulation study indicates that when the true effects cover the whole correlation block of the data, then the model with HL penalty has less sensitivity than L1. Only when the true effects are sparse as well, the sensitivity of HL penalty improves. This indication also translates to the aspect of prediction performance. Models with lasso and HL penalties have comparable prediction performance between them at their optimal λ. Only in the simulation study, the model with HL penalty has a slight advantage than that with lasso penalty when the true effects are very sparse.

Further improvement in prediction performance comes when we consider elastic net and HLnet penalties. The incorporation of ridge penalty improves considerably the sensitivity of HL penalty in the simulation study. In the real data, although the prediction performances of elastic net and HLnet models are the best, the HLnet penalty remains producing very sparse estimates while elastic net solution is closer to the ridge estimates. Therefore, the use of elastic net and HLnet in the prediction based on genomic data is recommended, with HLnet to be preferred when (very) sparse solution is desired.

Acknowledgements

The first author (ÖK) is supported by The Scientific and Technological Research Council of Turkey (TUBITAK), Grant Number: 1059B191601904, as part of 2219-International Postdoctoral Research Scholarship Programme.

References

[1] M.E. Ahsen, T.P. Boren, N.K. Singh, B. Misganaw, D.G. Mutch, K.N. Moore, F.J. Backes, C.K. McCourt, J.S. Lea, D.S. Miller, M.A. White, and M. Vidyasagar, *Sparse feature selection for classification and prediction of metastasis in endometrial cancer*, BMC Genomics 18 (2017), p. 233. Available at https://doi.org/10.1186/s12864-017-3604-y.

[2] I. Amarasena, S. Chatterjee, J. Walters, R. WoodBaker, and K. Fong, *Platinum versus nonplatinum chemotherapy regimens for small cell lung cancer*, Cochrane Database of Systematic Reviews (2015). Available at https://doi.org//10.1002/14651858.CD006849.pub3.

[3] C. Brennan, *et al.*, *The somatic genomic landscape of glioblastoma*, Cell 155 (2013), pp. 462 – 477. Available at http://www.sciencedirect.com/science/article/pii/S0092867413012087.

[4] S. Chapman, G. Robinson, J. Stradling, S. West, and J. Wrightson, *Oxford Handbook of Respiratory Medicine*, Oxford Medical Publications, Oxford University Press, 2014, Available at https://books.google.co.uk/books?id=hLqXAwAAQBAJ.

[5] G. Durif, L. Modolo, J. Michaelsson, J.E. Mold, S. Lambert-Lacroix, and F. Picard, *High dimensional classification with combined adaptive sparse PLS and logistic regression*, Bioinformatics 34 (2017), pp. 485–493. Available at https://dx.doi.org/10.1093/bioinformatics/btx571.

[6] B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani, *Least angle regression*, Ann. Statist. 32 (2004), pp. 407–499. Available at https://doi.org/10.1214/009053604000000067.

[7] M. Forero-Castro, C. Robledo, R. Benito, M. Abigar, A. frica Martn, M. Arefi, J.L. Fuster, N. de las Heras, J.N. Rodrguez, J. Quintero, S. Riesco, L. Hermosn, I. de la Fuente, I. Recio, J. Ribera, J. Labrador, J.M. Alonso, C. Olivier, M. Sierra, M. Megido, L.A. Corchete-Snchez, J. Ciudad Pizarro, J.L. Garca, J.M. Ribera, and J.M. Hernndez-

Rivas, *Genome-wide dna copy number analysis of acute lymphoblastic leukemia identifies new genetic markers associated with clinical outcome*, PLoS One 11 (2016), pp. 1–20. Available at https://doi.org/10.1371/journal.pone.0148972.

[8] A. Gusnanto, C.C. Taylor, I. Nafisah, H.M. Wood, P. Rabbitts, and S. Berri, *Estimating optimal window size for analysis of low-coverage next-generation sequence data*, Bioinformatics 30 (2014), pp. 1823–1829. Available at http://dx.doi.org/10.1093/bioinformatics/btu123.

[9] A. Gusnanto, H.M. Wood, Y. Pawitan, P. Rabbitts, and S. Berri, *Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data*, Bioinformatics 28 (2012), pp. 40–47. Available at http://dx.doi.org/10.1093/bioinformatics/btr593.

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

[11] J. Huang, A. Gusnanto, K. O'Sullivan, J. Staaf, . Borg, and Y. Pawitan, *Robust smooth segmentation approach for array cgh data analysis*, Bioinformatics 23 (2007), pp. 2463–2469. Available at http://dx.doi.org/10.1093/bioinformatics/btm359.

[12] J. Huang, A. Salim, K. Lei, K. O'Sullivan, and Y. Pawitan, *Classification of array cgh data using smoothed logistic regression model*, Statistics in Medicine 28 (2009), pp. 3798–3810. Available at https://onlinelibrary.wiley.com/doi/abs/10.1002/sim.3753.

[13] M. Jamal-Hanjani, G.A. Wilson, N. McGranahan, N.J. Birkbak, T.B. Watkins, S. Veeriah, S. Shafi, D.H. Johnson, R. Mitter, R. Rosenthal, M. Salm, S. Horswell, M. Escudero, N. Matthews, A. Rowan, T. Chambers, D.A. Moore, S. Turajlic, H. Xu, S.M. Lee, M.D. Forster, T. Ahmad, C.T. Hiley, C. Abbosh, M. Falzon, E. Borg, T. Marafioti, D. Lawrence, M. Hayward, S. Kolvekar, N. Panagiotopoulos, S.M. Janes, R. Thakrar, A. Ahmed, F. Blackhall, Y. Summers, R. Shah, L. Joseph, A.M. Quinn, P.A. Crosbie, B. Naidu, G. Middleton, G. Langman, S. Trotter, M. Nicolson, H. Remmen, K. Kerr, M. Chetty, L. Gomersall, D.A. Fennell, A. Nakas, S. Rathinam, G. Anand, S. Khan, P. Russell, V. Ezhil, B. Ismail, M. Irvin-Sellers, V. Prakash, J.F. Lester, M. Kornaszewska, R. Attanoos, H. Adams, H. Davies, S. Dentro, P. Taniere, B. OSullivan, H.L. Lowe, J.A. Hartley, N. Iles, H. Bell, Y. Ngai, J.A. Shaw, J. Herrero, Z. Szallasi, R.F. Schwarz, A. Stewart, S.A. Quezada, J. Le Quesne, P. Van Loo, C. Dive, A. Hackshaw, and C. Swanton, *Tracking the evolution of nonsmall-cell lung cancer*, New England Journal of Medicine 376 (2017), pp. 2109–2121. Available at https://doi.org/10.1056/NEJMoa1616288, PMID: 28445112.

[14] L. Jiang, Y. Xiao, Y. Ding, J. Tang, and F. Guo, *Discovering cancer subtypes via an accurate fusion strategy on multiple profile data*, Frontiers in Genetics 10 (2019), p. 20. Available at https://www.frontiersin.org/article/10.3389/fgene.2019.00020.

[15] Y. Lee, J. Nelder, and Y. Pawitan, *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood*, Chapman & Hall/CRC Monographs on Statistics & Applied Probability, CRC Press, 2006, Available at https://books.google.co.uk/books?id=JjDMeRM4woEC.

[16] Y. Lee and J.A. Nelder, *Hierarchical generalized linear models*, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1996), pp. 619–678.

[17] Y. Lee and H.S. Oh, *A new sparse variable selection via random-effect model*, Journal of Multivariate Analysis 125 (2014), pp. 89 – 99. Available at http://www.sciencedirect.com/science/article/pii/S0047259X13002583.

[18] H. Li and R. Durbin, *Fast and accurate short read alignment with burrowswheeler transform*, Bioinformatics 25 (2009), pp. 1754–1760. Available at http://dx.doi.org/10.1093/bioinformatics/btp324.

[19] Y. Pawitan, *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, 2013, Available at https://books.google.co.uk/books?id=8T8fAQAAQBAJ.

[20] L. Paz-Ares, A. Luft, D. Vicente, A. Tafreshi, M. Gumus, J. Mazieres, B. Hermes, F. Cay Senler, T. Csoszi, A. Fulop, J. Rodriguez-Cid, J. Wilson, S. Sugawara, T. Kato, K.H. Lee, Y. Cheng, S. Novello, B. Halmos, X. Li, G.M. Lubiniecki, B. Piperdi,

and D.M. Kowalski, *Pembrolizumab plus chemotherapy for squamous nonsmall-cell lung cancer*, New England Journal of Medicine 379 (2018), pp. 2040–2051. Available at https://doi.org/10.1056/NEJMoa1810865, PMID: 30280635.

[21] B. Rajaratnam, S. Roberts, D. Sparks, and O. Dalal, *Lasso regression: estimation and shrinkage via the limit of gibbs sampling*, Journal of the Royal Statistical Society: Series B (Statistical Methodology) 78 (2016), pp. 153–174. Available at https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/rssb.12106.

[22] R. Tibshirani, *Regression shrinkage and selection via the lasso*, Journal of the Royal Statistical Society. Series B (Methodological) 58 (1996), pp. 267–288. Available at http://www.jstor.org/stable/2346178.

[23] H.M. Wood, O. Belvedere, C. Conway, C. Daly, R. Chalkley, M. Bickerdike, C. McKinley, P. Egan, L. Ross, B. Hayward, J. Morgan, L. Davidson, K. MacLennan, T.K. Ong, K. Papagiannopoulos, I. Cook, D.J. Adams, G.R. Taylor, and P. Rabbitts, *Using next-generation sequencing for high resolution multiplex analysis of copy number variation from nanogram quantities of dna from formalin-fixed paraffin-embedded specimens*, Nucleic Acids Research 38 (2010), p. e151. Available at http://dx.doi.org/10.1093/nar/gkq510.

[24] T.T. Wu and K. Lange, *Coordinate descent algorithms for lasso penalized regression*, Ann. Appl. Stat. 2 (2008), pp. 224–244. Available at https://doi.org/10.1214/07-AOAS147.

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