



Deposited via The University of Leeds.

White Rose Research Online URL for this paper:

<https://eprints.whiterose.ac.uk/id/eprint/244993/>

Version: Accepted Version

Article:

Kaymaz, O., Vionanda, D., Dogru, F.Z. et al. (2026) Sparse Logistic Regression on Genomic Data for Prediction of Tumour Pathological Subtype. IEEE Transactions on Computational Biology and Bioinformatics. ISSN: 1545-5963

<https://doi.org/10.1109/TCBBIO.2026.3732668>

This is an author produced version of an article published in IEEE Transactions on Computational Biology and Bioinformatics, made available via the University of Leeds Research Outputs Policy under the terms of the Creative Commons Attribution License (CC-BY), which permits unrestricted use, distribution and reproduction in any medium, provided the original work is properly cited.

Reuse

This article is distributed under the terms of the Creative Commons Attribution (CC BY) licence. This licence allows you to distribute, remix, tweak, and build upon the work, even commercially, as long as you credit the authors for the original work. More information and the full terms of the licence here:

<https://creativecommons.org/licenses/>

Takedown

If you consider content in White Rose Research Online to be in breach of UK law, please notify us by emailing eprints@whiterose.ac.uk including the URL of the record and the reason for the withdrawal request.

Sparse logistic regression on genomic data for prediction of tumour pathological subtype

Özlem Kaymaz, Dodi Vionanda, Fatma Z. Dogru, Youngjo Lee, Henry M. Wood, and Arief Gusnanto

Abstract—The correct prediction of tumour subtype is critical for the treatment of cancer patients to maximise the chance of survival. The patients’ genomic information, such as copy number alterations (CNA) profile, has increasingly become an important factor in the prediction to supplement the traditional pathological subtyping. The incorporation of the CNA information in a prediction model, such as logistic regression, faces two major statistical challenges: first, how to estimate the model parameters in the thousands and, second, how to deal with the correlation of CNA between genomic regions. To address them, we propose a sparse logistic regression model with random effects where some of its parameters are estimated to zero while the other parameters are non-zero. In effect, a variable selection is embedded in the modelling. To deal with the correlation of CNA across genomic regions, we extend further the model to incorporate an additional penalty in the corresponding likelihood function in the logistic regression. The results show that we can identify selected genomic regions that are informative to distinguish different tumour subtypes, while giving a good prediction ability. We illustrate the methodology using CNA dataset from a lung cancer cohort.

Index Terms—Tumour, subtype, cancer, logistic regression, h -likelihood, prediction, copy number.

I. INTRODUCTION

IN the context of care and wellbeing of cancer patients, the prediction of tumour subtypes is critical because the treatment administered to the patients is dependent on their tumour subtypes [Amarasena *et al.* (2015)]. An incorrect prediction will likely result in either lowering the chance of a positive outcome or subjecting the patients to unnecessary pain. To increase the chance of a success treatment, prediction of tumour subtype based on the patients’ genomic information, such as copy number alterations (CNA), has been considered to supplement the traditional pathological subtyping. CNA are genomic alterations, in which some regions exhibit more or less than the normal two copies. CNA are informative for prediction of tumour subtype because each tumour subtype has different patterns of CNA. This is because the underlying process of tumour development is different between subtypes due to the origin of cancer [Jamal-Hanjani *et al.* (2017)].

Ö. Kaymaz is with the Department of Statistics, Ankara University, Ankara, Türkiye (Email: ozlgullu@ankara.edu.tr). D. Vionanda is with the Department of Statistics, Faculty of Mathematics and Natural Sciences, Padang State University, Padang 25171, Indonesia. F.Z. Dogru is with the Department of Statistics, Giresun University, Giresun, Türkiye. Y. Lee is with the Department of Statistics, Seoul National University, Republic of Korea. H.M. Wood is with the Leeds Institute of Medical Research at St. James’s and NIHR Leeds BRC, University of Leeds, Leeds LS9 7TF, United Kingdom. A. Gusnanto is with the Department of Statistics, School of Mathematics, University of Leeds, Leeds LS2 9JT, United Kingdom (Email: A.Gusnanto@leeds.ac.uk)

Manuscript received April 19, 2005; revised August 26, 2015.

A range of predictive models has been considered for tumour classification or risk stratification using copy number alteration (CNA) data. Machine-learning approaches such as deep neural networks [Islam *et al.* (2017)], [Karim *et al.* (2020)] and random forest [Gao and Baudis (2021)] have also been applied to CNA-derived features for tasks including molecular subtype classification. Logistic regression remains attractive because it provides class probabilities and interpretable regression coefficients [Elkhwagy *et al.* (2024)], and it has previously been used for subtype prediction from genome-wide CNA profiles. For example, [Shibahara *et al.* (2023)] use logistic regression method to identify intrinsic tumour subtypes from CNA of lung cancer patients. [Bareche *et al.* (2020)] consider logistic regression on breast-cancer patients’ genomic information to predict triple-negative breast cancer (TNBC) molecular subtypes. However, the adoption of logistic regression in our case faces some challenges as a result of the characteristics of the CNA data. First, it is well known that genome-wide CNA data are high dimensional and the number of genomic locations (predictors) far exceeds the number of patients. So, standard logistic regression model will be over-parameterised. Secondly, we expect that not all genomic locations are important for prediction. In other words, this is a variable selection challenge where we expect only CNA in some genomic locations that are relevant to distinguish the two subtypes, while CNA in the other regions are not necessarily relevant. For example, [Forero-Castro *et al.* (2016)] suggest that some genomic regions in chromosomes 1, 14, and 15 are important to distinguish leukaemia cancer subtypes. [Brennan *et al.* (2013)] also show that some genomic regions in chromosomes 4, 7, and 12 are relevant to distinguish subtypes in glioblastoma (a type of brain cancer). Thirdly, due to occurrence of CNA in segments, there are some spatial correlation of CNA along the genome. This can be attributed to the origin of CNA that involves non-allelic homologous recombination during meiotic recombination [Hastings *et al.* (2009)].

[Wang *et al.* (2022)] recently reviewed different sparse logistic regression models for prediction based on biological data, mainly incorporating lasso (L1) penalty [Tibshirani (1996)]. The lasso formulation [Tibshirani (1996)] has some drawbacks in dealing with correlated predictors as noted by [Zou and Hastie (2005)], who later extend the formulation to elastic-net model. In this study, we propose an alternative formulation of sparse logistic regression method to address the above challenges with random effects modelling framework [Lee and Nelder (1996)],

[Lee *et al.* (2006)]. Specifically, we propose a hierarchical model where the random effects are assumed to follow a normal distribution, but its variance parameter is assumed further to follow an inverse Gamma distribution. This approach produces a very sparse solution where some genomic locations have zero parameter estimates while the others have non-zero estimates, to address the first and second challenges. To deal with the third challenge of correlation between genomic locations, we extend further the model by assuming a mixture distribution for the random effects. The extension uses the same principle as the extension of lasso model [Tibshirani (1996)] into elastic-net model [Zou and Hastie (2005)]. In a simulation study, the proposed model has good operating characteristics. In general, the proposed model is recommended, especially when the CNA segments are small, or when the number of truly associated genomic regions are small.

The rest of the article is organised as follows. Section 2 describes the proposed methodology, including the datasets involved. Section 3 shows the results, and Section 4 presents the discussion. The conclusion is given in Section 5.

II. METHODS

A. Datasets

Our motivation comes from a study of lung cancer patients at the Leeds University Teaching Hospital (UK). Seventy six NSCLC patients are considered for this study. Out of these 76 patients, 38 patients have adenocarcinoma tumour subtype and the other 38 patients have squamous carcinoma subtype. The details on biological sample preparation, DNA extraction and sequence preparation are described in [Wood *et al.* (2010)]. The DNA samples are sequenced using Illumina GAIIX sequencer to produce short DNA sequences ('reads'). The short sequences are aligned to the assembly hg19 of the human reference genome using the software package bwa suite version 0.5.9-r16 [Li and Durbin (2009)]. For our analysis, we only consider reads that can be uniquely mapped to the human reference genome with quality score ≥ 37 .

To estimate the CNA for each patient, we consider 'depth of coverage' by counting the number of reads in each fixed-size non-overlapping genomic region ('window'). The size of this genomic window is calculated using the *NGSoptwin* package [Gusnanto *et al.* (2014)] to be 150 kb. As a result, the whole genome can be covered by slightly more than 20,000 genomic windows. By excluding some missing values (e.g. the centromere regions), we consider 17,571 genomic windows for our analysis. The CNA profile for each patient is then obtained as a set of normalised quantitative estimates across genomic windows, using the *CNAnorm* package [Gusnanto *et al.* (2012)] to account for contamination of normal cells. An example of CNA estimates for a patient is given in Figure 1 (top panel). The figure indicates the CNA estimates for Patient LS63 in the study and highlights the spatial dependencies between genomic regions. This is better seen in Figure 1 (bottom panel) that describes the (Pearson) correlation of CNA between genomic windows in chromosome 5 across 76 patients in our study. The correlation between genomic regions is not only generally high, but also exhibits 'block' correlation between neighbouring genomic regions.

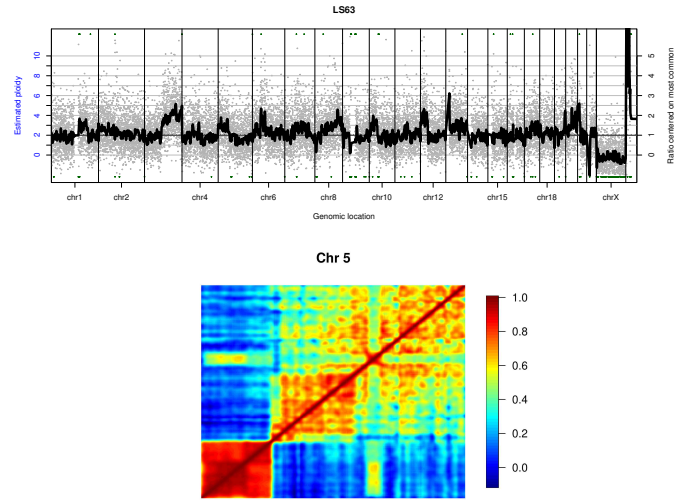


Fig. 1. Top panel: Segmented copy number alteration (CNA) data (solid line) for patient LS63 across genomic locations. Vertical lines separate different chromosomes, while grey dots correspond to the normalised copy number ratio. Bottom panel: Correlation of CNA between genomic windows in chromosome 5 across all patients.

To illustrate the generalisability of the proposed methodology to independent cohorts and a different representation of CNA data, we additionally analyse The Cancer Genome Atlas (TCGA) [Weinstein *et al.* (2013)] datasets via the University of California, Santa Cruz (UCSC) Xena public data portal [Goldman *et al.* (2020)], which hosts harmonized datasets generated by the Genomic Data Commons (GDC)[Grossman *et al.* (2016)]. We consider the somatic copy number data for oesophageal carcinoma (ESCA), lung squamous cell carcinoma (LUSC), and lung adenocarcinoma (LUAD). Copy number profiles were generated from Affymetrix Genome-Wide Human SNP Array 6.0 (SNP6) data on matched tumour-normal samples and processed through the GDC copy number variation pipeline based on the ASCAT framework (ASCAT3)[Ross *et al.* (2021)]. Further details of the data preprocessing steps are available from the National Cancer Institute website [NCI GDC (2026)]. For TCGA ESCA dataset, we have 56,974 gene-level copy numbers from 73 patients with squamous carcinoma subtype and 57 patients with adenocarcinoma subtype. For the TCGA lung cancer dataset, we have 56,433 gene-level copy numbers from 250 patients with squamous carcinoma subtype and 278 patients with adenocarcinoma. In these TCGA datasets, our main interest is still in the prediction of squamous/adeno carcinoma subtypes. Analyses of the TCGA data are included as independent applications demonstrating the generalisability of the methodology across different CNA representations, rather than as external validation of the Leeds data analysis. The analysis results of the TCGA datasets are mainly presented in the Supplementary Material.

Importantly, the predictors used in our models are quantitative normalised CNA estimates at the genomic-window or gene

level, rather than merely binary gain/loss indicators. Hence, the magnitude (or intensity) of copy number change is retained in the input through the estimated CNA values for each genomic region. We did not additionally construct alternative summary features such as cohort-level CNA frequency, because our focus in this paper is subject-specific prediction directly from each patient's genome-wide CNA profile once comparable CNA estimates have been obtained. A systematic investigation of the impact of alternative preprocessing choices, including different normalisation or segmentation procedures, would be valuable but is beyond the scope of the present methodological study.

B. Sparse logistic regression

In this section we describe the logistic regression model that we consider. This approach is still within the context of generalised linear models with random effects [Lee and Nelder (1996)], [Lee *et al.* (2006)]. Before we outline the proposed model, we denote first the notation that we shall use here. Let y_i be the i -th patient's tumour subtype, either as squamous carcinoma (coded as '1') or adenocarcinoma (coded as '0'), where $i = 1, 2, \dots, n$. We assume that the random variable associated with the i -th patient's tumour subtype is assumed to follow a Binomial($1, \mu_i$) distribution. Denoting $\mu \equiv (\mu_1, \mu_2, \dots, \mu_n)$, we model the mean vector μ as

$$h(\mu) = X\beta + Zb, \quad (1)$$

where $h(\cdot)$ is a logit link function that applies element-wise, X is a matrix of (fixed) predictors of size $n \times p$, β is a p -vector associated with the predictors in X , Z is a matrix of size $n \times q$ of CNA from those n patients, where q is the number of genomic windows in each patient, and b is a q -vector of random effects. In our application, X here is taken as a matrix of (clinical) predictors that contain a vector of ones for a (fixed) intercept.

The random effects b in many general applications are assumed to follow a normal distribution with mean zero and some variance, i.e. $b \sim N(0, D(\theta))$, where $D(\theta) \equiv \theta I_q$, θ is a variance parameter (or tuning parameter), and I_q is an identity matrix of size q . This formulation of b is considered as the ridge model [Kawaguchi *et al.* (2022)]. However, it is well known that the ridge model cannot produce a sparse solution. The term *sparse* here refers to some model parameters that are estimated as zero while the other parameters are estimated to be non-zero. Therefore, variable selection is not possible with this ridge model formulation, which makes it non-applicable in our case.

To achieve a sparse solution, the lasso model was previously proposed by [Tibshirani (1996)]. A study by [Zou and Hastie (2005)], however, indicates that this approach is problematic when the predictors are correlated, such as in our case. In particular, they note that the lasso model will select randomly a predictor to have non-zero estimate in a correlation block. The authors [Zou and Hastie (2005)] propose an elastic-net model where the lasso model is extended by including a ridge penalty function. They show that the ridge penalty function has a "grouping effect" on the estimate, which

helps the interpretation when the predictors are correlated. In this study, we propose a new alternative model to produce a sparse solution and extend it further in the same principle as the extension of lasso model to the elastic net model.

To proceed with the proposed model formulation, we assume that the random effects b in Equation (1) to follow a normal distribution such that

$$b_j | u_j \sim N(0, u_j \theta) \quad (2)$$

where θ represents the tuning parameter, $\theta = 2\sigma^2$, and u_j is assumed further to follow an inverse gamma (IG) distribution with parameters α and δ and mean one. This assumption indicates the hierarchical structure of the random effects in the modelling, and a novel approach to the original formulation of [Lee and Oh (2014)] and [Kaymaz *et al.* (2020)] that previously used an assumption of gamma distribution on u_j . We find that the use of inverse gamma, instead of gamma distribution, produces a more stable computation in practice, which we will revisit in the discussion. Under the inverse gamma distribution, the probability density function (pdf) of u_j is therefore given by

$$p(u_j; \alpha, \delta) = \frac{\delta^\alpha}{\Gamma(\alpha)} u_j^{-\alpha-1} \exp\left(-\frac{\delta}{u_j}\right), \quad (3)$$

where Γ is the gamma function, $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} \exp(-t) dt$. The mean of u_j is given by $E(u_j) = \frac{\delta}{\alpha-1}$ and, since we set $E(u_j) = 1$, the above pdf of u_j can be rewritten as

$$p(u_j; \alpha) = \frac{(\alpha-1)^\alpha}{\Gamma(\alpha)} u_j^{-\alpha-1} \exp\left(-\frac{(\alpha-1)}{u_j}\right). \quad (4)$$

Now, α is the parameter that determines the distribution of $p(u_j; \alpha)$ as will be described in Section II-C.

Within the framework of hierarchical likelihood (h -likelihood) [Lee and Nelder (1996)], [Lee *et al.* (2006)] and following [Lee and Oh (2014)], we can express the Equation (2) as

$$b_j = \sqrt{\tau_j} e. \quad (5)$$

Here, τ_j equals $2\sigma^2 u_j$ and e follows the standard normal distribution ($e \sim N(0, 1)$). Using the log-link, it can be written as

$$\log(\tau_j) = \log(2\sigma^2 + v_j), \quad (6)$$

where $v_j = \log(u_j)$.

The joint log likelihood of (β, θ, b) of the logistic regression model in Equation (1) can be expressed as [Lee and Nelder (1996)], [Lee *et al.* (2006)]

$$\ell(\beta, \theta, b) = \log p(y|b) + \log p(b). \quad (7)$$

The term $\log p(y|b)$ is the log likelihood function of the logistic regression model

$$\log p(y|b) = \sum_{i=1}^n \{y_i \log \mu_i + (1 - y_i) \log (1 - \mu_i)\}, \quad (8)$$

where

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

and X_i and Z_i are the i -th row of X and Z , respectively. Furthermore, $\log p(b)$ gives the hierarchical log likelihood of the random effects expressed by Equation (2) such that

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

where

$$\log p(b_j|u_j) = -\frac{1}{2} \left\{ \log(4\pi\sigma^2) + \log u_j + \frac{b_j^2}{2\sigma^2 u_j} \right\} \quad (10)$$

and

$$\begin{aligned} \log p(u_j) &= \alpha \log(\alpha - 1) - \log \Gamma(\alpha) \\ &\quad - (\alpha + 1) \log u_j - (\alpha - 1)/u_j. \end{aligned} \quad (11)$$

C. Parameter estimation

We now describe the estimation of β , b , and θ . In principle, the estimation of β and b is done on a fixed θ using an iterative weighted least squares (IWLS) algorithm. The estimation of θ is done separately by identifying the value of θ that minimises the Akaike's information criterion (AIC) across different values of θ as described below. It is important to note that, in the above formulation, α is set in advance (fixed) and it is not a parameter to estimate. At a fixed θ , the estimation of β and b is done by maximising the profile h -likelihood function [Lee and Nelder (1996)], [Lee *et al.* (2006)], [Lee and Oh (2014)]

$$\ell_p(\beta, \theta, b) = \log p(y|b)|_{u=\hat{u}} + \log p(b)|_{u=\hat{u}}. \quad (12)$$

To obtain $\hat{u}_j$, we identify the solution of $\partial \ell_p / \partial u = 0$ [Lee and Oh (2014)] and this gives

$$\hat{u}_j = \hat{u}_j(b_j) = \frac{1}{(2\alpha + 3)} \left(\frac{b_j^2}{2\sigma^2} + 2(\alpha - 1) \right). \quad (13)$$

In Equation (13), $\hat{u}_j$ is only a function of b_j because this calculation is performed at a given θ ($\theta = 2\sigma^2$), in addition to a given α .

The choice of α determines the type of solution for $\hat{b}$. In Equation (13), $\hat{u}_j$ goes to one as α goes to infinity ($\alpha \rightarrow \infty$) in limit terms. Therefore, the penalty function $\log p(b)$ in Equation (7) is proportional to that on the ridge model, or

$$\log p(b) \propto \sum_{j=1}^q b_j^2. \quad (14)$$

On the other hand, when α goes to one (from the right) in limit terms, the penalty function $\log p(b)$ in Equation (7) produces a sparse solution on $\hat{b}$. This can be visualised better in Figure 2. The shape of the penalty function $\log p(b)$ for $\alpha = 10$ already resembles that of a ridge penalty (i.e. quadratic shape). When $\alpha = 1.1$, the shape of $\log p(b)$ starts put more penalty to $\hat{b}$ than the ridge penalty around zero. When α is very close to one (in our case we use $\alpha = 1.00001$), the penalty function $\log p(b)$ produces a sparse solution for $\hat{b}$ because the function immediately penalise severely the estimates in the vicinity of zero. With this shape of $\log p(b)$, we expect that many estimates will be penalised to zero while still allowing some

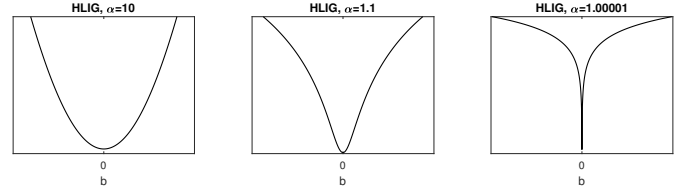


Fig. 2. The shape of penalty function $\log p(b)$ for different α values

estimates to be estimated away from zero. This is the type of solution we aim to achieve and, for practical purposes, we call this as hierarchical likelihood with inverse gamma (HLIG) model.

To proceed with the estimation of β and b using IWLS, we denote $\lambda \equiv \frac{1}{\theta}$ and denote $W \equiv \text{diag}(\frac{1}{\hat{u}_j})$ as a diagonal matrix of size $q \times q$ whose (diagonal) elements are the inverse of $\hat{u}_j$. Then, the mixed model equations [Lee and Nelder (1996)], [Lee *et al.* (2006)], [Pawitan (2013)] can be obtained by differentiating $\ell_p(\beta, \theta, b)$ in Equation (12) with respect to β and b at fixed θ

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

Here, Σ is an $n \times n$ diagonal matrix with diagonal elements $\Sigma_{ii} = \mu_i(1 - \mu_i)$ and Y is a working vector whose elements are given by

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

Therefore, β and b can be estimated iteratively by solving the following equations simultaneously (at fixed λ)

$$\hat{\beta} = (X^T \Sigma^{-1} X)^{-1} X^T \Sigma^{-1} (Y - Z \hat{b}), \quad (15)$$

$$\hat{b} = (Z^T \Sigma^{-1} Z + \lambda W)^{-1} Z^T \Sigma^{-1} (Y - X \hat{\beta}). \quad (16)$$

To estimate $\lambda \equiv \frac{1}{\theta}$, we calculate $AIC = -2 \log p(y|b) + 2df$ where $df = \text{trace}\{(Z^T \Sigma^{-1} Z + \lambda W)^{-1} Z^T \Sigma^{-1} Z\}$ for different values of λ and identify λ that minimises AIC [Pawitan (2013)].

D. Extension to HLIGnet model

As mentioned above, the HLIG model is expected to produce a sparse estimate of $\hat{b}$. However, to deal with the correlation of CNA between neighbouring genomic windows, we need to extend the model in the same principle as that of [Zou and Hastie (2005)] when they extended the lasso model [Tibshirani (1996)] to elastic-net model. The elastic-net model was introduced to deal with the sparse estimate produced by the lasso model when dealing with correlated predictors by adding an additional ridge penalty function to the lasso penalty function. In our study, we also extend the HLIG penalty function by adding the ridge penalty function. The addition of the ridge penalty within the HLIG modelling framework is possible because the ridge penalty function can be achieved by setting $\alpha \rightarrow \infty$ (or, "large enough") and the sparse HLIG solution is achieved when $\alpha \rightarrow 1$ in limit terms (in practice, we set $\alpha = 1.00001$). So, both penalty functions can be

expressed within the HLIG modelling framework. We denote the resulting penalty function as HLIGnet model.

To proceed, the addition of the ridge penalty function is done by assuming that the random effects b follow a mixture distribution with a density function

$$p_{\text{HLIGnet}}(b) = C(\theta, c)p(b; u_{\alpha_1}, \theta)^c p(b; u_{\alpha_2}, \theta)^{1-c} \quad (17)$$

where $C(\theta, c)$ is the normalising constant, c is the mixture proportion $0 \leq c \leq 1$, α_1 is set to one in limit terms (HLIG penalty function), and α_2 is set 'large enough' ($\alpha_2 \rightarrow \infty$, ridge penalty function). It is important to note that $C(\theta, c)$ does not depend on b . This is important when we wish to identify $\hat{u}$ as described below.

The log likelihood of the HLIGnet model can therefore be defined as

$$\ell_{\text{HLIGnet}}(\beta, \theta, b) = \log p(y|b) + \log p_{\text{HLIGnet}}(b), \quad (18)$$

where $\log p(y|b)$ is the log likelihood of the logistic regression model in Equation (8). The hierarchical log likelihood on the random effects can be written as

$$\begin{aligned} \log p_{\text{HLIGnet}}(b) &= cp_{\alpha_1}(b) + (1-c)p_{\alpha_2}(b) + \log C(\theta, c) \\ &= \log C(\theta, c) \\ &\quad + c \left\{ \sum_{j=1}^q \log p(b_j | u_{j,\alpha_1}) + \sum_{j=1}^q \log p(u_{j,\alpha_1}) \right\} \\ &\quad + (1-c) \left\{ \sum_{j=1}^q \log p(b_j | u_{j,\alpha_2}) \right. \\ &\quad \left. + \sum_{j=1}^q \log p(u_{j,\alpha_2}) \right\} \quad (19) \end{aligned}$$

where

$$\begin{aligned} \log p(b_j | u_{j,\alpha_1}) &= -\frac{1}{2} \left\{ \log(4\pi\sigma^2) + \log p(u_{j,\alpha_1}) + \frac{b_j^2}{2\sigma^2 u_{j,\alpha_1}} \right\}, \\ \log p(b_j | u_{j,\alpha_2}) &= -\frac{1}{2} \left\{ \log(4\pi\sigma^2) + \log p(u_{j,\alpha_2}) + \frac{b_j^2}{2\sigma^2 u_{j,\alpha_2}} \right\}, \\ \log p(u_{j,\alpha_1}) &= \alpha_1 \log(\alpha_1 - 1) - \log \Gamma(\alpha_1) \\ &\quad - (\alpha_1 + 1) \log(u_{j,\alpha_1}) - \frac{(\alpha_1 - 1)}{u_{j,\alpha_1}}, \\ \log p(u_{j,\alpha_2}) &= \alpha_2 \log(\alpha_2 - 1) - \log \Gamma(\alpha_2) \\ &\quad - (\alpha_2 + 1) \log(u_{j,\alpha_2}) - \frac{(\alpha_2 - 1)}{u_{j,\alpha_2}}. \end{aligned}$$

Given $(\alpha_1, \alpha_2, \theta)$, the estimation of β and b is done using the profile h -likelihood [Lee *et al.* (2006)], [Lee and Nelder (1996)], [Lee and Oh (2014)]

$$\ell_{p,\text{HLIGnet}}(\beta, \theta, p) = \log p(y|b) \Big|_{u=\hat{u}} + \log p_{\text{HLIGnet}}(b) \Big|_{u=\hat{u}}. \quad (20)$$

Following [Lee and Oh (2014)], we obtain $\hat{u}_j$ as a solution to u ($\partial \ell_{p,\text{HLIGnet}} / \partial u = 0$) and this yields

$$\hat{u}_j = \hat{u}(b_j) \approx c \hat{u}(b_j, \alpha_1) \Big|_{\alpha_1 \rightarrow 1} + (1-c) \hat{u}(b_j, \alpha_2) \Big|_{\alpha_2 \rightarrow \infty} \quad (21)$$

where

$$\hat{u}(b_j, \alpha_1) = \frac{b_j^2}{10\sigma^2}, \quad \hat{u}(b_j, \alpha_2) = 1 \quad (22)$$

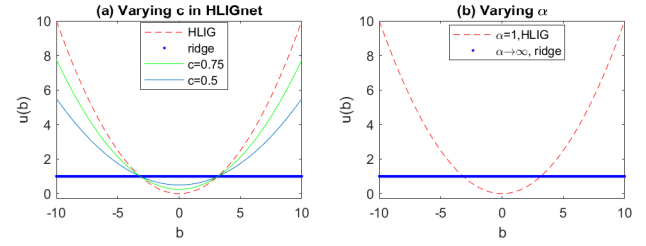


Fig. 3. The shape of u_j as a function of b_j , given a fixed θ (or, equivalently σ^2) for (a) different values of c in the HLIGnet and (b) different values of α 's.

Figure 3 shows the shape of $\hat{u}_j$ as a function of b_j . Note that in the IWLS, the matrix W is a diagonal matrix, whose diagonal elements are $1/\hat{u}_j$, for $j = 1, 2, \dots, q$. Hence, the lower $\hat{u}_j$ is, the more penalty that is imposed on b_j in its estimation. In Figure 3(a), for different values of mixing proportion c , the figure shows that the HLIGnet model transforms into the ridge model when $c = 0$ and the HLIGnet model will be the HLIG model when $c = 1$. In the ridge solution, the fact that $u_j = 1$ for different values of b_j indicates that the penalty is applied in equal amount. In the HLIG model solution, there is a huge penalty imposed around zero, and still 'allow' the estimates of b_j to be away from zero. In the implementation of IWLS, if $\hat{u}_j$ is zero, we add a small constant (10^{-8}) to $\hat{u}_j$ to avoid singularity. As a result, the estimate $\hat{b}_j$ cannot be exactly zero in the iteration. Therefore, when $\hat{b}_j$ is small enough (e.g. 10^{-5}) in IWLS, we set the estimate to zero. Figure 3(b) indicates the shape of function of u_j as a function of b_j for $\alpha = 1$ and $\alpha \rightarrow \infty$. As we have described above, we can infer that the former gives HLIG sparse solution while the latter gives the ridge solution.

Simulation study

To understand the proposed models' working characteristics, we perform a simulation study. In principle, we first simulate CNA datasets, including the (simulated) tumour pathological subtype. Subsequently, we fit different models to the simulated data, including the proposed models, as described further below. From this process, we are able to identify, first, the sensitivity and specificity of the random effects estimates and, second, the prediction error using cross-validation, across different simulation settings (the terms *sensitivity* and *specificity* in our context are described below).

The CNA data are generated as follows. A matrix of copy number ratios [Gusnanto *et al.* (2012)], denoted Z^* of size $n = 100 \times 1000 = q$ are generated as $Z^* \sim \text{MVN}(\mathbf{0}, \Psi)$ where MVN is the multivariate normal density function and Ψ is 1000×1000 covariance matrix. For a known K' , the covariance matrix is defined as

$$\Psi = \begin{bmatrix} \Psi_{11} & \Psi_{12} & 0 & \cdots & 0 \\ \Psi_{21} & \Psi_{22} & \Psi_{23} & \cdots & 0 \\ 0 & \Psi_{32} & \Psi_{33} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & \Psi_{K'K'} \end{bmatrix}. \quad (23)$$

This indicates that the variables of Z^* (corresponding to genomic regions) are block structured. For the diagonal blocks, Ψ_{kk} is of size $K \times K$, $K = (1000/K')$, $k = 1, 2, \dots, K'$, 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 $\rho = 0.9$. This high correlation within a diagonal block reflects the case we observe in real data. For the one-off diagonal blocks, we define $\Psi_{k(k+1)} = \Psi_{(k+1)k}$ of size $K \times K$, $k = 1, 2, \dots, K' - 1$, as

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

We consider two different values of r in this simulation study: $r = 0$ and $r = 0.4$. The first case indicates that the covariance matrix Ψ is a block diagonal matrix with independent blocks. The second case indicates that Ψ is a block diagonal matrix with correlated blocks. The latter setting with $r = 0.4$ reflects better the characteristics that we observe in the real data (Figure 1).

After the matrix of copy number ratios Z^* is generated, we perform a circular binary segmentation [Olshen *et al.* (2004)] on each row of Z^* (corresponding to each patient). The segmented lines, as a result of the segmentation, are in essence the simulated copy number alterations and summarised in a CNA matrix Z (still of size $n \times q$).

For the purpose of this simulation study, we consider three different values of K : 10, 20, and 40. This corresponds to different numbers of correlation block in Ψ : 100, 50, and 25 (respectively). To simulate the true tumour pathological subtype for patient i , $i = 1, 2, \dots, n$, we generate the mean μ_i as

$$\mu_i = \frac{\exp\{Z_i b\}}{1 + \exp\{Z_i b\}}$$

and the subtype y_i is generated as Binomial(1, μ_i) distribution. The true random effects $b \equiv \{b_j\}$ here are generated according to 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

$$b = (\underbrace{1, 1, \dots, 1}_K, \underbrace{-1, -1, \dots, -1}_K, 0, 0, \dots, 0)^T.$$

- **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

$$b = (\underbrace{1, 0, \dots, 0}_K, \underbrace{-1, 0, \dots, 0}_K, 0, 0, \dots, 0)^T.$$

In scenario A, the truly associated variables are those in the first two segments (correlation blocks in Ψ) and take the whole

segment. In scenario B, the truly associated variables are only a ‘singleton’ in each of the first two segments of correlated variables in Z .

Once the simulated data are created, we fit four different models: lasso (L1), HLIG, elastic net (enet), and HLIGnet. The lasso and elastic-net models are included as the main benchmark methods because they are the most established sparse logistic regression approaches for high-dimensional data, with the elastic net specifically designed to improve over lasso in the presence of correlated predictors [Tibshirani (1996)], [Zou and Hastie (2005)]. This makes them the most directly relevant comparators for the proposed HLIG and HLIGnet models, which are constructed within the same logistic-regression regularisation framework. Each model is fitted to simulated data under each of combination of settings: (1) the size of the segments or correlation blocks (K) is 10, 20, and 40, (2) the configuration of truly associated variables (scenarios A and B above), and (3) the status of the diagonal blocks in the covariance matrix Ψ (whether they are independent or correlated). Overall, there are 12 dataset settings for each model and we generate independently 100 simulated datasets for each simulation setting. Each model is fitted to each simulated data and we calculate (1) the classification error of prediction from a cross validation, and (2) *sensitivity* and *specificity* of sparse solution across the different configurations while keeping the overall computational burden manageable. It is important to note that the sensitivity and specificity reported here refer to the recovery of the sparse solution, rather than to the classification of tumour subtype. Specifically, we define *sensitivity* as the proportion of truly non-zero random effects (b_j 's) that are correctly estimated as non-zero, and *specificity* as the proportion of truly zero random effects that are correctly estimated as zero. These two quantities are therefore measures of variable-selection performance, and are useful because our interest is not only in prediction but also in identifying which genomic regions are retained in the model. For predictive performance itself, we report the cross-validation classification error and the area under the ROC curve (AUC), although the latter is presented in the Supplementary Material.

It is important to note that some settings of the simulation study are much more difficult to address than that we find in real data. This is a deliberate attempt to challenge the proposed methods close to their practical limits in predicting tumour pathological subtype while at the same time identifying important genomic regions. For example, we have a very high level of correlation within the correlation block. Furthermore, the two different scenarios (A and B) represent two different extremes: 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. Identifying the one true effect within the correlation block in scenario B will be extremely difficult since the other variables within the block are equally good.

III. RESULTS

A. Simulation data

The results of the simulation study in terms of prediction error (from cross validation) are presented in Figures 4 and

5. Figure 4 shows the prediction errors of different models across 100 simulated datasets generated under independent correlation blocks. Overall, the figure indicates that the prediction error under the HLIG model is comparable to that under L1 model, and the prediction error of HLIGnet model is always better than that under Enet model. Under scenario A, the prediction error for HLIGnet model is always higher than those of L1 and HLIG models, although, in scenario B, the prediction error for HLIGnet model is better than or at least comparable to those of L1 and HLIG models.

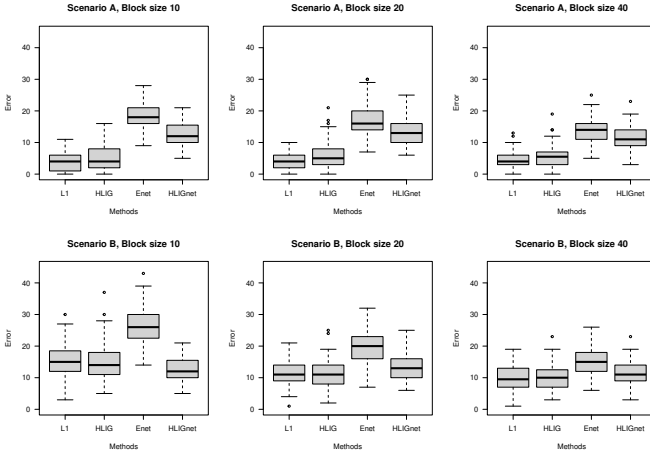


Fig. 4. Classification error (in percent) of logistic regression models with lasso penalty (L1), HLIG penalty (HLIG), elastic net penalty (Enet), and HLIGnet penalty (HLIGnet) from 100 simulated data under different simulation settings when the diagonal blocks of the data covariance matrix are independent ($r = 0$). The results for correlated diagonal blocks are presented in Figure 5. Top row is for scenario A, and bottom row is for scenario B. Left to right columns correspond to increasing block size. Further details of the simulation settings are described in the main text Section II-E.

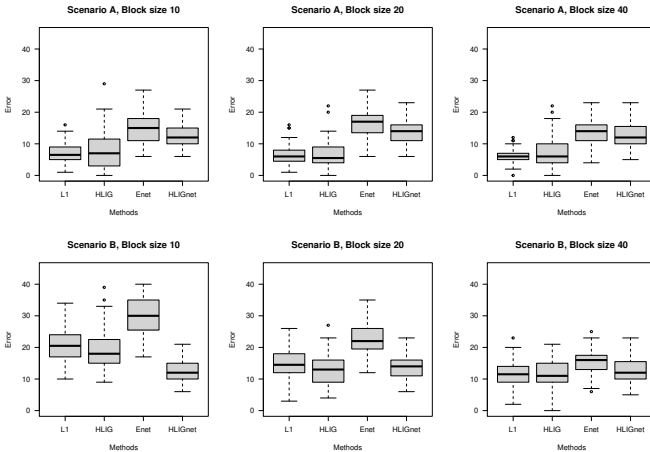


Fig. 5. Classification error (in percent) of logistic regression models with lasso penalty (L1), HLIG penalty (HLIG), elastic net penalty (Enet), and HLIGnet penalty (HLIGnet) from 100 simulated data under different simulation settings when the diagonal blocks of the data covariance matrix are moderately correlated ($r = 0.4$). The results for independent diagonal blocks are presented in Figure 4. Top row is for scenario A, and bottom row is for scenario B. Left to right columns correspond to increasing block size. Further details of the simulation settings are described in the main text Section II-E.

Figure 5 shows the prediction errors of different models

across 100 simulated datasets generated under correlated correlation blocks. For scenario A, Figure 5 indicates that the prediction error under the HLIG model is comparable to that under L1 model, and the prediction error of HLIGnet model is always better than that under Enet model. Furthermore, the prediction error for HLIGnet model is always higher than those of L1 and HLIG models. This is similar to the pattern seen in Figure 4. In scenario B, however, the prediction error for HLIG and HLIGnet models is better than L1 and Enet models when the block size is low. When the block size is high, the prediction error for both models is comparable to that of L1 model. Indeed, in scenario B, the prediction error of the models considered in this simulation study tends to be comparable when the correlation block size is high. It is important to note that the simulation setting of correlated diagonal blocks in the covariance matrix is closer to the situation that we observe in the real data.

In our context, classification error remains the primary predictive metric because the practical objective is to minimise incorrect subtype assignment. We also report the AUC as a complementary threshold-independent measure of discrimination in the Supplementary Material. However, AUC should be interpreted with caution, as it summarises performance over all possible thresholds and does not directly reflect the error rate at the decision threshold used in practice [Verbakel *et al.* (2020)], [Janssens and Martens(2020)], [Hand(2009)].

Furthermore, Table I shows the sensitivity and specificity of the random effects estimates for L1, HLIG, Enet, and HLIGnet under different simulation settings. The table indicates that the L1 and HLIG models have relatively high specificity, but low or moderate sensitivity, with the sensitivity of HLIG model is lower than that of L1 model. This is as expected because HLIG model produces much sparse estimates than L1 model does: The HLIG penalty function puts much faster and greater penalty than that of L1 penalty as the parameters are estimated away from zero (see Figure 2). Furthermore, the sensitivity for Enet and HLIGnet models is in general very high with moderate specificity. The inclusion of the ridge penalty in the both models has a ‘grouping effect’ as previously noted by [Zou and Hastie (2005)]. In the solution, the random effects estimates from correlated variables tend to group together. This is the main reason for the high sensitivity. However, this comes at a price on the specificity, as it goes down from that in the L1 and HLIG models.

Although the sensitivity for the HLIG model is lower than that of L1 model, this does not necessarily translate into worse prediction as shown in Figures 4 and 5. Similarly, the high sensitivity of the Enet model does not necessarily translate into better prediction. In fact, both figures indicate that the Enet model has the worst prediction error in a cross validation. As noted above, the HLIGnet model is more promising in general in the context of correlated correlation block (Figure 5) where it strikes a good balance between sensitivity, specificity, and prediction error. It is important to note that the low sensitivity of the HLIG model in this simulation shows the limit of the methodology when dealing with such a challenging simulation setting: other variables are considered equally good

TABLE I

SENSITIVITY AND SPECIFICITY (WITH STANDARD ERROR) OF THE DIFFERENT LOGISTIC REGRESSION MODELS WITH LASSO PENALTY (L1), HLIG PENALTY (HLIG), ELASTIC NET PENALTY (ENET), AND HLIGNET PENALTY (HLIGNET) FROM 100 SIMULATED DATA ACROSS DIFFERENT SIMULATION SETTINGS.

Block Size	Methods (penalty)	Independent block			
		Scenario A		Scenario B	
		Sens.	Spec.	Sens.	Spec.
10	L1	0.624 (0.019)	0.969 (0.001)	0.535 (0.015)	0.890 (0.003)
	HLIG	0.110 (0.005)	0.999 (0.000)	0.105 (0.020)	0.993 (0.000)
	Enet	1.000 (0.000)	0.374 (0.006)	1.000 (0.000)	0.292 (0.005)
	HLIGNet	0.978 (0.008)	0.511 (0.009)	0.750 (0.025)	0.410 (0.007)
20	L1	0.470 (0.017)	0.962 (0.002)	0.425 (0.018)	0.903 (0.003)
	HLIG	0.053 (0.003)	0.999 (0.000)	0.070 (0.017)	0.995 (0.000)
	Enet	1.000 (0.000)	0.370 (0.007)	1.000 (0.000)	0.294 (0.006)
	HLIGNet	0.962 (0.011)	0.469 (0.009)	0.730 (0.025)	0.383 (0.008)
40	L1	0.383 (0.019)	0.957 (0.003)	0.400 (0.020)	0.895 (0.003)
	HLIG	0.078 (0.002)	0.999 (0.000)	0.075 (0.018)	0.996 (0.000)
	Enet	0.998 (0.001)	0.328 (0.010)	1.000 (0.000)	0.315 (0.009)
	HLIGNet	0.938 (0.014)	0.441 (0.012)	0.815 (0.024)	0.354 (0.009)
Block Size	Methods (penalty)	Correlated block			
		Scenario A		Scenario B	
		Sens.	Spec.	Sens.	Spec.
10	L1	0.641 (0.019)	0.953 (0.002)	0.500 (0.000)	0.893 (0.002)
	HLIG	0.106 (0.005)	0.999 (0.000)	0.085 (0.019)	0.992 (0.000)
	Enet	1.000 (0.000)	0.386 (0.009)	1.000 (0.000)	0.244 (0.005)
	HLIGNet	0.954 (0.012)	0.492 (0.009)	0.725 (0.025)	0.339 (0.006)
20	L1	0.470 (0.021)	0.951 (0.002)	0.450 (0.017)	0.910 (0.002)
	HLIG	0.057 (0.005)	0.999 (0.000)	0.050 (0.015)	0.994 (0.000)
	Enet	1.000 (0.000)	0.313 (0.007)	1.000 (0.000)	0.239 (0.005)
	HLIGNet	0.943 (0.012)	0.410 (0.008)	0.715 (0.025)	0.309 (0.007)
40	L1	0.373 (0.020)	0.953 (0.002)	0.425 (0.018)	0.899 (0.003)
	HLIG	0.027 (0.002)	0.999 (0.000)	0.060 (0.016)	0.994 (0.000)
	Enet	0.997 (0.001)	0.264 (0.009)	0.995 (0.005)	0.224 (0.007)
	HLIGNet	0.907 (0.017)	0.375 (0.010)	0.730 (0.025)	0.264 (0.008)

Figures in brackets are their respective standard errors. All figures are rounded to three decimal digits. We define *sensitivity* as the proportion of truly non-zero random effects (b_j 's) that are correctly estimated as non-zero, and *specificity* as the proportion of truly zero random effects that are correctly estimated as zero. Further details are on the simulation settings are described in the main text of Section II-E.

for prediction due to their correlation. Nevertheless, the HLIG model is still able to achieve comparable prediction error to that of the L1 model, even with the ‘wrong’ variables in the model.

B. Lung cancer data

We now present the results of fitting the proposed models to the lung cancer data.

TABLE II

THE ESTIMATES (FIXED EFFECTS), STANDARD ERRORS (*Std. Err.*), *z*-VALUES, AND STATISTICAL SIGNIFICANCE OF *Age* AND *Sex* FROM HLIG AND HLIGNET MODEL FITTINGS IN THE LUNG CANCER DATASET.

Term	Estimate	Std. Err.	<i>z</i> -value	<i>p</i> -value
<i>HLIG model</i>				
Intercept	-8.408	14.155	-0.594	0.552
Age	0.182	0.190	0.957	0.338
Sex	-3.487	2.374	-1.469	0.142
<i>HLIGNet model</i>				
Intercept	-12.802	36.287	-0.352	0.724
Age	0.058	0.411	0.141	0.888
Sex	-1.717	5.554	-0.309	0.757

Table II presents the estimates of the fixed effects ($\hat{\beta}$) and their significance. The table indicates that *Age* and *Sex* are not significant in the lung cancer data to distinguish the two subtypes in both HLIG and HLIGNet models. In a larger population-based study, we usually identify a difference in the subtype incidence among males and females [Guarga *et al.* (2021)]. With limited number of patients in our study, the power to identify these differences is arguably too low.

The estimates of the random effects ($\hat{b}$) for HLIG and HLIGNet models are presented in Figure 6. As a comparison in terms of the type of solution produced, the figure also shows the estimates under the lasso (L1) and elastic-net models [Zou and Hastie (2005)]. For the HLIG model, there are only four non-zero estimates and 17,567 zero estimates out of 17,571 predictors (genomic windows). They correspond to genomic windows in Chromosomes 3, 14, 17, and 20. As a comparison, the lasso (L1) model produced 112 non-zero estimates (approximately 0.64%) and 17,459 zero estimates. They correspond to genomic windows in Chromosomes 3, 5, 9, 14, 17, and 20. The non-zero estimates under HLIG model are either also estimated as non-zero under the lasso model, or have a non-zero estimate nearby. Moreover, the sign (of the random effect estimates) under the HLIG model is also consistent the corresponding estimates under the lasso model.

Figure 6 also presents the random effects estimates for HLIGNet model. The figure shows that there is one predictor whose estimate is relatively large (relative to the rest of the estimates). When we look into the range of majority of the estimates (“zoom in”), the figure for HLIGNet model indicates that there are 9,746 zero estimates (approximately 55.5%) and 7,825 non-zero estimates (approximately 44.5%). This is in contrast to the figure for the Elastic-net model, in which there are only 26 genomic windows with zero estimates (approximately 0.15%). The figure for the Elastic-net model suggests that the estimates are closer towards ridge model solution than lasso model solution. It is important to note that the non-zero estimates of the HLIGNet exhibit ‘grouping effect’ [Zou and Hastie (2005)] because of the addition of the ridge penalty to the HLIG penalty. In fact, there are 248 groups of non-zero estimates. The definition of ‘group’ here is cluster of consecutive non-zero random-effects estimates (along the

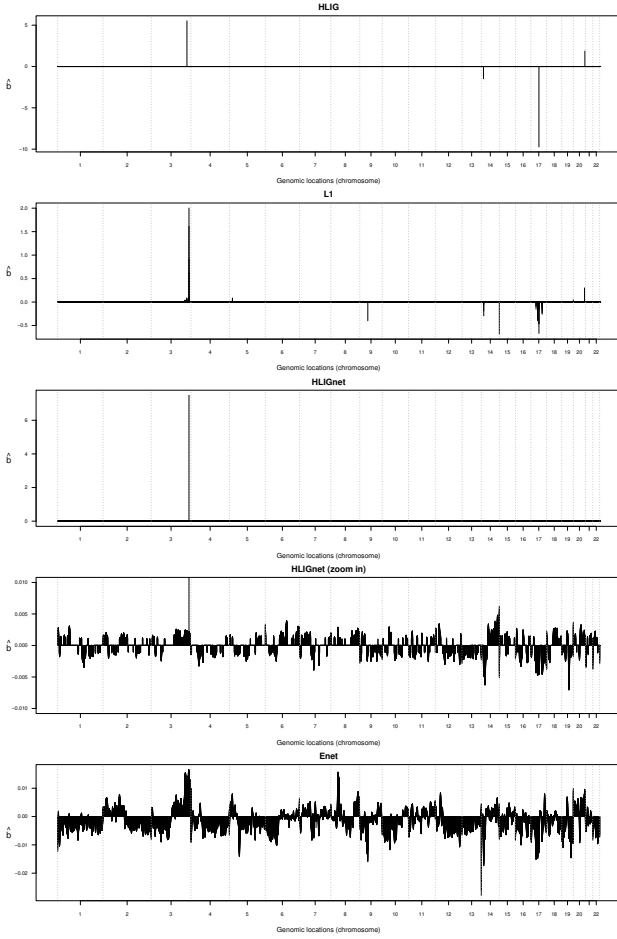


Fig. 6. Estimates of the random effects ($\hat{b}$) from logistic regression for HLIG, lasso (L1), HLIGnet, and elastic-net (Enet) models.

genome). It is interesting to note that the number of predictors non-zero estimates in HLIGnet model (7,825) is very close to the number of genomic regions to differ between the two groups identified by powerful two-sample tests as noted by [Telford *et al.* (2020)] who previously worked on the same data. They are 7,981, 7,868, and 7,558 genomic regions for Cramer test, Anderson-Darling test, and Cramer-von Mises test, respectively.

Because one of the practical motivations for sparse solution is biological interpretability, we further examined the selected genes from the TCGA oesophageal cancer and lung cancer analyses because both datasets already contain CNA information at the gene level. A focused functional interpretation of the regions identified by HLIG and HLIGnet in the TCGA datasets is provided in the Supplementary Material. These annotations are intended to illustrate the biological plausibility of the selected features and to highlight candidate regions for follow-up investigation, rather than to claim that they are already validated stand-alone biomarkers.

In terms of predictive ability of those models, we consider a five-fold cross-validation to calculate the classification error across different values of tuning parameter λ . Although one can use this as a method to identify optimal λ (as an alternative to AIC above), the purpose here is simply to understand the

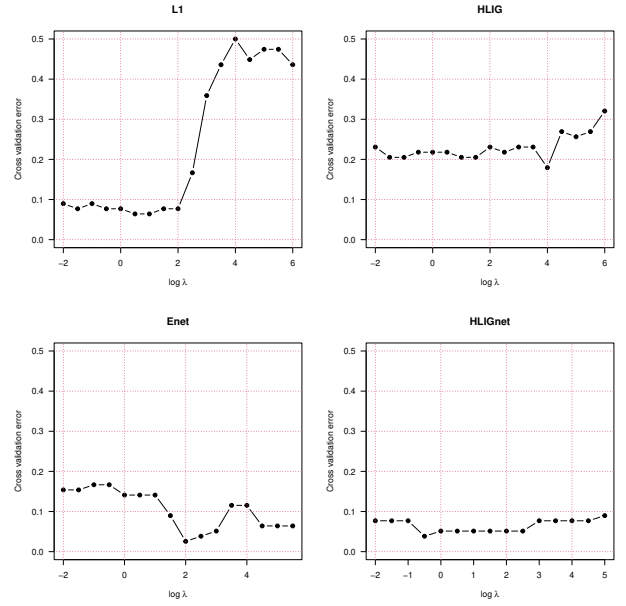


Fig. 7. Cross-validation classification error from logistic regression for lasso (L1), HLIG, elastic-net (Enet), and HLIGnet models on the lung cancer data (across different values of λ).

characteristics of prediction error across the tuning parameter. For example, is the predictive ability of the models is sensitive to the value of λ ? The results of the cross-validation are presented in Figure 7. The figure indicates that both the lasso (L1) and HLIG models tend to have relatively higher classification error than those of elastic-net (Enet) and HLIGnet. The error of lasso model is relatively low for low λ and increase substantially after a point. In comparison, the error of HLIG model tend to stay moderate across λ . The classification errors for the elastic-net and models are much improved compared to those of the lasso (L1) and the HLIG models. This indicates that the addition of the ridge penalty to the lasso and HLIG models managed to address our concerns on the correlation between predictors (genomic regions), although we observed in Figure 6 that the solution produced by the elastic net model falls short to produce sufficiently sparse solution. From a practical perspective, Figure 7 highlights an important trade-off between predictive accuracy, tuning robustness, and interpretability. The lasso model attains relatively low classification error only over a limited range of small λ , after which the error increases markedly, suggesting that its predictive performance may be sensitive to the choice of tuning parameter. In contrast, the HLIG model shows a more stable error profile across λ , but its error remains higher overall, indicating that the gain in sparsity may come at the expense of prediction. The elastic-net and HLIGnet models achieve consistently lower classification errors, which is practically important when the aim is to reduce subtype misclassification. However, Figure 6 shows that the elastic-net solution is only weakly sparse, whereas HLIGnet still retains substantial sparsity and grouping of neighbouring genomic regions. Therefore, for routine application, HLIGnet appears to offer the most useful compromise: it improves predictive performance while preserving a level of model

parsimony that aids biological interpretation.

IV. DISCUSSION

We have investigated logistic regression models with random effects for prediction of tumour subtypes in the context of genomic CNA data. The proposed formulation enables a sparse solution while performing prediction with good operating characteristics. [Zou and Hastie (2005)] have previously noted that lasso model has some drawbacks in dealing with correlated data. Our study supports this assessment and our results indicate that both lasso and HLIg models are less than optimal when dealing with correlated data. Only after the inclusion of ridge penalty in the model to create elastic-net [Zou and Hastie (2005)] and HLIgnet models, respectively, the performance of the models improves. We find that the elastic-net model produces a less-than-satisfactory sparse solution as only a small proportion of the predictors are zero estimated. The HLIgnet model, on the other hand, strike a good balance between sparse solution, ‘grouping effect’ [Zou and Hastie (2005)], and a good prediction. We recommend to consider the HLIgnet model for routine analysis, if prediction performance and interpretation are the main considerations.

In terms of methodology, the use of inverse gamma distribution in the model hierarchy has similar results to the original use of gamma distribution [Lee and Oh (2014)], [Kaymaz *et al.* (2020)] in terms of creating a sparse solution and an extension to deal with spatial correlation. We view this novel approach as an alternative formulation, and it demonstrates the generalisability and flexibility of the h -likelihood formulation [Lee and Nelder (1996)], [Lee *et al.* (2006)]. Although the original gamma distribution assumption gives a stable computation for a range of tuning parameters, our experience suggested that the use of inverse gamma distribution in the assumption gives a stable computation for a wider range of tuning parameters and starting values. This is an advantage in practical application when we consider different datasets with different tuning parameters.

Although we illustrate the application of the methodology in the context of CNA data, it is also generally applicable to other datasets with high-dimensional predictors and exhibit ‘block’ correlation. For example, logistic regression models have been considered for genome-wide DNA methylation analysis (see e.g. [Sun and Wang (2012)], [Choi *et al.* (2018)]) where the CpG islands have created correlations. Logistic regressions models have also been considered for (genome-wide) genetic association analysis on single nucleotide polymorphisms (see e.g. [Ayers and Cordell (2010)], [Austin *et al.* (2013)], [Prive *et al.* (2019)]) that exhibit block correlations due to linkage disequilibrium. The TCGA analyses demonstrate the generalisability of the methodology to gene-level copy-number data. Because the predictor definitions and genomic resolutions differ, they do not constitute formal external validation of the Leeds data analysis. We have adopted an efficient method in the computation to scale up to higher dimensional data by ensuring that matrix inversion involves an $n \times n$ matrix, and optimised for vectors in a multi-thread environment. In the

TCGA lung cancer datasets with almost 57 thousand predictors and 528 patients, the HLIg and HLIgnet model fittings take up 113 and 71 seconds per value of λ , respectively, using eight cores in the computation. In the TCGA oesophageal cancer with similar number of predictors and 130 patients, the HLIg and HLIgnet model fittings take up 27 and 17 seconds per value of λ , respectively.

The present study focuses on binary subtype prediction within a logistic-regression framework. An extension to multicategory outcomes would be a natural next step, for example through multinomial or cumulative-logit formulations in the broader generalised linear modelling context. Likewise, richer clinical covariates, such as tumour stage, anatomical location, or tumour size, can in principle be incorporated through the fixed-effects component whenever these variables are available with sufficient completeness. In the datasets considered here, however, clinical annotation was limited: the Leeds cohort only recorded age and sex/gender consistently, and in TCGA the inclusion of additional clinical variables would have resulted in a substantial reduction in analysable sample size because of missing data. We therefore restricted the present analyses to the covariates available with adequate completeness, while emphasising that this is a practical limitation of the current datasets rather than of the proposed model formulation.

An additional advantage of the proposed sparse formulations is that they facilitate biological follow-up by reducing a very large set of genomic predictors to a smaller number of interesting genes. In the Supplementary Material, we provide focused functional annotations for genes identified in the TCGA oesophageal cancer and lung cancer analyses, in order to illustrate the type of biological interpretation enabled by the sparse solution. We emphasise, however, that the present study is methodological in scope: the identified genes should be viewed as hypothesis-generating candidates whose biological significance and reproducibility require further evaluation in independent datasets and, ultimately, experimental validation. In this context, reproducibility may be assessed at several levels, including recurrence of selected regions across related datasets, concordance between HLIg and HLIgnet selections, and consistency with known cancer-related pathways or previously reported genomic features.

A further potential extension of the proposed framework is multi-omics integration. In principle, the current formulation can be extended by including additional omics-specific design matrices, each with its own vector of random effects, alongside the CNA matrix. For example, copy number predictors could be analysed jointly with DNA methylation, SNP, or other genomic-level measurements within the same generalised linear modelling framework. Such an extension would be attractive because multi-omics integration has been shown in the broader literature to improve cancer subtype classification and to provide a more comprehensive view of tumour biology [Qi *et al.* (2021)], [Guo *et al.* (2023)], [Kim *et al.* (2023)], [Itai *et al.* (2023)]. At the same time, this would introduce additional methodological issues, including cross-platform normalisation, missingness across omics layers, and the choice of regularisation structure across different data blocks. For this reason, we regard multi-omics

HLIG/HLIGnet modelling as a promising direction for future work rather than an objective of the present study. A broad range of classification methods was previously evaluated using the Leeds CNA data, with logistic regression providing the best cross-validated prediction among the methods considered [Gusnanto *et al.* (2015)]. The present study therefore focuses on the remaining methodological challenge of combining prediction with sparse recovery, biological interpretation, and fixed-effect inference within the logistic-regression framework.

Lastly, we recognise that other machine learning methods, including random forests, boosting methods, and deep learning, may also be considered for prediction from genomic data. However, these approaches address a somewhat different modelling objective from that of the present study. In high-dimensional genomic settings, random forests are known to face difficulties in variable-importance assessment and feature selection when predictors are strongly correlated, while deep learning methods often require additional representation learning, feature extraction, or data-balancing strategies to cope with small-sample, high-dimensional inputs and can be less transparent in biological interpretation [Strobl *et al.* (2007)], [Darst *et al.* (2018)], [Gregorutti *et al.* (2017)], [Miotto *et al.* (2018)], [Shen *et al.* (2022)], [Liu *et al.* (2022)]. Likewise, boosting-based CNA prediction models in the literature are often built on engineered summary features such as arm-level copy-number changes, breakpoints, or CNA signatures, rather than directly on the original genome-wide predictor set [Yao *et al.* (2023)], [Yu *et al.* (2020)].

In contrast, the purpose of HLIg and HLIgnet is to remain within a sparse logistic-regression framework so that prediction, sparsity, grouping of correlated regions, and coefficient-based interpretation can be studied together in a common model. This also allows clinical covariates to be incorporated through the fixed-effects component and, in our setting, enables standard inferential summaries for those fixed effects. For these reasons, a full comparison with deep learning or boosting approaches would require a separate investigation with dedicated preprocessing, feature-engineering, and tuning pipelines, and is therefore beyond the scope of the present methodological paper.

V. CONCLUSION

This study has suggested that the proposed HLIgnet model is recommended for prediction and interpretation of tumour subtypes based on copy number alterations data. The method is expected to be generalisable for other types of omics data, where correlated high-dimensional predictors are common. The HLIgnet model strikes a good balance between sparsity of the solution that aids the interpretation, and the predictive ability.

ACKNOWLEDGMENT

Henry M. Wood and Arief Gusnanto are supported in part by the National Institute for Health and Care Research (NIHR) Leeds Biomedical Research Centre (BRC) (NIHR203331). The views expressed are those of the author(s) and not

necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

REFERENCES

- [Amarasena *et al.* (2015)] Amarasena IU, Chatterjee S, Walters JAE, Wood-Baker R, Fong KM. Platinum versus non-platinum chemotherapy regimens for small cell lung cancer. *Cochrane Database of Systematic Reviews* 2015, Issue 8. Art. No.: CD006849. DOI: 10.1002/14651858.CD006849.pub3. Accessed 01 August 2024.
- [Austin *et al.* (2013)] Austin E, Pan W, Shen X. Penalized Regression and Risk Prediction in Genome-Wide Association Studies. *Stat Anal Data Min.* 2013 Aug 1;6(4):10.1002/sam.11183. doi: 10.1002/sam.11183. PMID: 24348893; PMCID: PMC3859439.
- [Ayers and Cordell (2010)] Ayers KL, Cordell HJ. SNP selection in genome-wide and candidate gene studies via penalized logistic regression. *Genet Epidemiol.* 2010 Dec;34(8):879-91. doi: 10.1002/gepi.20543. PMID: 21104890; PMCID: PMC3410531.
- [Bareche *et al.* (2020)] Bareche Y, Buisseret L, Gruosso T, Girard E, Venet D, Dupont F, Desmedt C, Lamsimont D, Park M, Rothé F, Stagg J, Sotiriou C. Unraveling Triple-Negative Breast Cancer Tumor Microenvironment Heterogeneity: Towards an Optimized Treatment Approach. *J Natl Cancer Inst.* 2020 Jul 1;112(7):708-719. doi: 10.1093/jnci/djz208. PMID: 31665482; PMCID: PMC7357326.
- [Brennan *et al.* (2013)] Brennan CW, Verhaak RG, McKenna A, Campos B, Nounshmeir H, Salama SR, Zheng S, Chakravarty D, Sanborn JZ, Berman SH, Beroukhi R, Bernard B, Wu CJ, Genovese G, Shmulevich I, Barnholtz-Sloan J, Zou L, Vegesna R, Shukla SA, Ciriello G, Yung WK, Zhang W, Sougnez C, Mikkelsen T, Aldape K, Bigner DD, Van Meir EG, Prados M, Sloan A, Black KL, Eschbacher J, Finocchiaro G, Friedman W, Andrews DW, Guha A, Iacocca M, O'Neill BP, Foltz G, Myers J, Weisenberger DJ, Penny R, Kucherlapati R, Perou CM, Hayes DN, Gibbs R, Marra M, Mills GB, Lander E, Spellman P, Wilson R, Sander C, Weinstein J, Meyerson M, Gabriel S, Laird PW, Haussler D, Getz G, Chin L; TCGA Research Network. The somatic genomic landscape of glioblastoma. *Cell.* 2013 Oct 10;155(2):462-77. doi: 10.1016/j.cell.2013.09.034. Erratum in: *Cell.* 2014 Apr 24;157(3):753. PMID: 24120142; PMCID: PMC3910500.
- [Choi *et al.* (2018)] Choi J, Kim K, Sun H. New variable selection strategy for analysis of high-dimensional DNA methylation data. *J Bioinform Comput Biol.* 2018 Aug;16(4):1850010. doi: 10.1142/S0219720018500105. Epub 2018 Apr 26. PMID: 29954287.
- [Darst *et al.* (2018)] Darst BF, Malecki KC, Engelman CD. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data. *BMC Genet.* 2018 Sep 17;19(Suppl 1):65. doi: 10.1186/s12863-018-0633-8. PMID: 30255764; PMCID: PMC6157185.
- [Elkhwaghy *et al.* (2024)] Elkhwaghy, D.M.A.S., Kiriacos, C.J. & Mansour, M. Logistic regression and other statistical tools in diagnostic biomarker studies. *Clin Transl Oncol* 26, 2172–2180 (2024). <https://doi.org/10.1007/s12094-024-03413-8>
- [Forero-Castro *et al.* (2016)] Forero-Castro M, Robledo C, Benito R, Abáigar M, África Martín A, Arefi M, Fuster JL, de Las Heras N, Rodríguez JN, Quintero J, Riesco S, Hermosín L, de la Fuente I, Recio I, Ribera J, Labrador J, Alonso JM, Olivier C, Sierra M, Megido M, Corchete-Sánchez LA, Ciudad Pizarro J, García JL, Ribera JM, Hernández-Rivas JM. Genome-Wide DNA Copy Number Analysis of Acute Lymphoblastic Leukemia Identifies New Genetic Markers Associated with Clinical Outcome. *PLoS One.* 2016 Feb 12;11(2):e0148972. doi: 10.1371/journal.pone.0148972. PMID: 26872047; PMCID: PMC4752220.
- [Gao and Baudis (2021)] Gao B and Baudis M (2021) Signatures of Discriminative Copy Number Aberrations in 31 Cancer Subtypes. *Front. Genet.* 12:654887. doi: 10.3389/fgene.2021.654887
- [Goldman *et al.* (2020)] Goldman MJ, Craft B, Hastie M, Repecka K, McDade F, Kamath A, Banerjee A, Luo Y, Rogers D, Brooks AN, Zhu J, Haussler D. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat Biotechnol.* 2020 Jun;38(6):675-678.
- [Gregorutti *et al.* (2017)] Gregorutti B, Michel B, Saint-Pierre P. Correlation and variable importance in random forests. *Statistics and Computing.* 2017;27:659–678. doi:10.1007/s11222-016-9646-1
- [Grossman *et al.* (2016)] Grossman RL, Heath AP, Ferretti V, Varmus HE, Lowy DR, Kibbe WA, Staudt LM. Toward a Shared Vision for Cancer Genomic Data. *N Engl J Med.* 2016 Sep 22;375(12):1109-12

- [Guarga *et al.* (2021)] Guarga L, Ameijide A, Marcos-Gragera R, Carulla M, Delgadillo J, Borràs JM, Galceran J. Trends in lung cancer incidence by age, sex and histology from 2012 to 2025 in Catalonia (Spain). *Sci Rep.* 2021 Dec 2;11(1):23274. doi: 10.1038/s41598-021-02582-8. PMID: 34857781; PMCID: PMC8639747.
- [Guo *et al.* (2023)] Guo H, Lv X, Li Y, Li M. Attention-based GCN integrates multi-omics data for breast cancer subtype classification and patient-specific gene marker identification. *Briefings in Functional Genomics.* 2023;22(5):463–474. doi:10.1093/bfgp/elad013
- [Gusnanto *et al.* (2014)] Gusnanto A, Taylor CC, Nafisah I, Wood HM, Rabbitts P, Berri S. Estimating optimal window size for analysis of low-coverage next-generation sequence data. *Bioinformatics.* 2014 Jul 1;30(13):1823–9. doi: 10.1093/bioinformatics/btu123. Epub 2014 Mar 5. PMID: 24603986.
- [Gusnanto *et al.* (2012)] Gusnanto A, Wood HM, Pawitan Y, Rabbitts P, Berri S. Correcting for cancer genome size and tumour cell content enables better estimation of copy number alterations from next-generation sequence data. *Bioinformatics.* 2012 Jan 1;28(1):40–7. doi: 10.1093/bioinformatics/btr593. Epub 2011 Oct 28. PMID: 22039209.
- [Gusnanto *et al.* (2015)] Gusnanto A, Tcherveniakov P, Shuweihi F, Samman M, Rabbitts P, Wood HM. Stratifying tumour subtypes based on copy number alteration profiles using next-generation sequence data. *Bioinformatics.* 2015;31(16):2713–2720. doi:10.1093/bioinformatics/btv191
- [Hand (2009)] Hand DJ. (2009) Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning.* 77:103–123. doi:10.1007/s10994-009-5119-5
- [Hastings *et al.* (2009)] Hastings PJ, Lupski JR, Rosenberg SM, Ira G. Mechanisms of change in gene copy number. *Nat Rev Genet.* 2009 Aug;10(8):551–64. doi: 10.1038/nrg2593. PMID: 19597530; PMCID: PMC2864001.
- [Islam *et al.* (2017)] Islam, M.M., Ajwad, R., Chi, C., Domaratzki, M., Wang, Y., Hu, P. (2017). Somatic Copy Number Alteration-Based Prediction of Molecular Subtypes of Breast Cancer Using Deep Learning Model. In: Mouhoub, M., Langlais, P. (eds) *Advances in Artificial Intelligence. Canadian AI 2017. Lecture Notes in Computer Science()*, vol 10233. Springer, Cham. https://doi.org/10.1007/978-3-319-57351-9_7
- [Itai *et al.* (2023)] Itai Y, Rappoport N, Shamir R. Integration of gene expression and DNA methylation data across different experiments. *Nucleic Acids Research.* 2023;51(15):7762–7776. doi:10.1093/nar/gkad566
- [Jamal-Hanjani *et al.* (2017)] Jamal-Hanjani M, Wilson GA, McGranahan N, Birkbak NJ, Watkins TBK, Veeriah S, Shafi S, Johnson DH, Mitter R, Rosenthal R, Salm M, Horswell S, Escudero M, Matthews N, Rowan A, Chambers T, Moore DA, Turajlic S, Xu H, Lee SM, Forster MD, Ahmad T, Hiley CT, Abbosh C, Falzon M, Borg E, Marafioti T, Lawrence D, Hayward M, Kolvekar S, Panagiotopoulos N, Janes SM, Thakrar R, Ahmed A, Blackhall F, Summers Y, Shah R, Joseph L, Quinn AM, Crosbie PA, Naidu B, Middleton G, Langman G, Trotter S, Nicolson M, Remmen H, Kerr K, Chetty M, Gomersall L, Fennell DA, Nakas A, Rathinam S, Anand G, Khan S, Russell P, Ezhil V, Ismail B, Irvin-Sellers M, Prakash V, Lester JF, Kornaszewska M, Attanoos R, Adams H, Davies H, Dentro S, Taniere P, O'Sullivan B, Lowe HL, Hartley JA, Iles N, Bell H, Ngai Y, Shaw JA, Herrero J, Szallasi Z, Schwarz RF, Stewart A, Quezada SA, Le Quesne J, Van Loo P, Dive C, Hackshaw A, Swanton C; TRACERx Consortium. Tracking the Evolution of Non-Small-Cell Lung Cancer. *N Engl J Med.* 2017 Jun 1;376(22):2109–2121. doi: 10.1056/NEJMoa1616288. Epub 2017 Apr 26. PMID: 28445112.
- [Janssens and Martens (2020)] Janssens ACJW, Martens FK. Reflection on modern methods: Revisiting the area under the ROC Curve. *Int J Epidemiol.* 2020 Aug 1;49(4):1397–1403. doi: 10.1093/ije/dy274.
- [Karim *et al.* (2020)] Karim, M.R., Rahman, A., Jares, J.B. et al. A snapshot neural ensemble method for cancer-type prediction based on copy number variations. *Neural Comput & Applic* 32, 15281–15299 (2020). <https://doi.org/10.1007/s00521-019-04616-9>
- [Kawaguchi *et al.* (2022)] Kawaguchi ES, Li S, Weaver GM, Lewinger JP. Hierarchical Ridge Regression for Incorporating Prior Information in Genomic Studies. *J Data Sci.* 2022 Jan;20(1):34–50. doi: 10.6339/21-jds1030. Epub 2021 Dec 13. PMID: 36274755; PMCID: PMC9581069.
- [Kaymaz *et al.* (2020)] Kaymaz Ö, Alqahtani K, Wood HM, Gusnanto A. Prediction of tumour pathological subtype from genomic profile using sparse logistic regression with random effects. *J Appl Stat.* 2020 Mar 11;48(4):605–622.
- [Kim *et al.* (2023)] Kim SY, Jeong S, Lee W, Jeon Y, Kim Y-J, Park S, Lee D, Go D, Song S-H, Lee S, Woo HG, Yoon J-K, Park YS, Kim YT, Lee S-H, Kim KH, Lim Y, Kim J-S, Kim H-P, Bang D, Kim T-Y. Cancer signature ensemble integrating cDNA methylation, copy number, and fragmentation facilitates multi-cancer early detection. *Experimental & Molecular Medicine.* 2023;55:2445–2460. doi:10.1038/s12276-023-01119-5
- [Lee and Nelder (1996)] Y. Lee, J. A. Nelder, Hierarchical Generalized Linear Models, *Journal of the Royal Statistical Society: Series B (Methodological)*, Volume 58, Issue 4, November 1996, Pages 619–656. <https://doi.org/10.1111/j.2517-6161.1996.tb02105.x>
- [Lee *et al.* (2006)] Lee, Y., Nelder, J.A., & Pawitan, Y. (2006). *Generalized Linear Models with Random Effects: Unified Analysis via H-likelihood* (1st ed.). Chapman and Hall/CRC. <https://doi.org/10.1201/9781420011340>
- [Lee and Oh (2014)] Lee, Youngjo & Oh, Hee-Seok, 2014. A new sparse variable selection via random-effect model, *Journal of Multivariate Analysis*, 125, pages 89–99.
- [Li and Durbin (2009)] Li H, Durbin R. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics (Oxford, England).* 2009 Jul;25(14):1754–1760. DOI: 10.1093/bioinformatics/btp324. PMID: 19451168; PMCID: PMC2705234.
- [Liu *et al.* (2022)] Liu L, Meng Q, Weng C, Lu Q, Wang T, Wen Y. Explainable deep transfer learning model for disease risk prediction using high-dimensional genomic data. *PLoS Computational Biology.* 2022;18(7):e1010328. doi:10.1371/journal.pcbi.1010328
- [Miotto *et al.* (2018)] Miotto R, Wang F, Wang S, Jiang X, Dudley JT. Deep learning for healthcare: review, opportunities and challenges. *Briefings in Bioinformatics.* 2018;19(6):1236–1246. doi:10.1093/bib/bbx044
- [NCI GDC (2026)] National Cancer Institute Genomic Data Commons, Copy Number Variation Analysis Pipeline, https://docs.gdc.cancer.gov/Data/Bioinformatics_Pipelines/CNV_Pipeline/, accessed 30 March 2026.
- [Olshen *et al.* (2004)] Olshen AB, Venkatraman ES, Lucito R, Wigler M. Circular binary segmentation for the analysis of array-based DNA copy number data. *Biostatistics.* 2004 Oct;5(4):557–72. doi: 10.1093/biostatistics/kxh008. PMID: 15475419.
- [Pawitan (2013)] Pawitan, Y (2013) *In All Likelihood: Statistical Modelling and Inference Using Likelihood*, Oxford University Press, New York
- [Prive *et al.* (2019)] Prive F, Aschard H, Blum MGB. Efficient Implementation of Penalized Regression for Genetic Risk Prediction. *Genetics.* 2019 May;212(1):65–74. doi: 10.1534/genetics.119.302019. Epub 2019 Feb 26. PMID: 30808621; PMCID: PMC6499521.
- [Qi *et al.* (2021)] Qi L, Wang W, Wu T, Zhu L, He L, Wang X. Multi-Omics Data Fusion for Cancer Molecular Subtyping Using Sparse Canonical Correlation Analysis. *Frontiers in Genetics.* 2021;12:607817. doi:10.3389/fgene.2021.607817
- [Ross *et al.* (2021)] Ross EM, Haase K, Van Loo P, Markowitz F. Allele-specific multi-sample copy number segmentation in ASCAT. *Bioinformatics.* 2021 Jul 27;37(13):1909–1911.
- [Shibahara *et al.* (2023)] Shibahara T, Wada C, Yamashita Y, Fujita K, Sato M, Kuwata J, et al. (2023) Deep learning generates custom-made logistic regression models for explaining how breast cancer subtypes are classified. *PLoS ONE* 18(5): e0286072. <https://doi.org/10.1371/journal.pone.0286072>
- [Shen *et al.* (2022)] Shen J, Shi J, Luo J, Zhai H, Liu X, Wu Z, Yan C, Luo H. Deep learning approach for cancer subtype classification using high-dimensional gene expression data. *BMC Bioinformatics.* 2022;23(1):430. doi:10.1186/s12859-022-04980-9
- [Strobl *et al.* (2007)] Strobl C, Boulesteix AL, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. *BMC Bioinformatics.* 2007 Jan 25;8:25. doi: 10.1186/1471-2105-8-25. PMID: 17254353; PMCID: PMC1796903.
- [Sun and Wang (2012)] Sun H, Wang S. Penalized logistic regression for high-dimensional DNA methylation data with case-control studies. *Bioinformatics.* 2012 May 15;28(10):1368–75. doi: 10.1093/bioinformatics/bts145. Epub 2012 Mar 30. PMID: 22467913; PMCID: PMC3348559.
- [Telford *et al.* (2020)] Telford A., Taylor C. C., Wood H. M., Wood H. C., Gusnanto A. Properties and approximate p-value calculation of the Cramer test. *Journal of Statistical Computation and Simulation.* 2020. Vol. 90. No. 11. pp. 1965–1981.
- [Tibshirani (1996)] Tibshirani, R. (1996). Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society: Series B (Methodological)*, 58(1), 267–288.
- [Verbakel *et al.* (2020)] Verbakel JY, Steyerberg EW, Uno H, De Cock B, Wynants L, Collins GS, Van Calster B. ROC curves for clinical prediction models part 1. ROC plots showed no added value above the AUC when evaluating the performance of clinical prediction models. *J Clin Epidemiol.* 2020 Oct;126:207–216. doi: 10.1016/j.jclinepi.2020.01.028.
- [Wang *et al.* (2022)] Wang P, Chen S, Yang S. Recent Advances on Penalized Regression Models for Biological Data. *Mathematics.* 2022; 10(19):3695. <https://doi.org/10.3390/math10193695>

- [Weinstein *et al.* (2013)] Weinstein JN, Collisson EA, Mills GB, Shaw KR, Ozenberger BA, Ellrott K, Shmulevich I, Sander C, Stuart JM. The Cancer Genome Atlas Pan-Cancer analysis project. *Nat Genet.* 2013 Oct;45(10):1113-20
- [Wood *et al.* (2010)] Wood HM, Belvedere O, Conway C, Daly C, Chalkley R, Bickerdike M, McKinley C, Egan P, Ross L, Hayward B, Morgan J, Davidson L, MacLennan K, Ong TK, Papagiannopoulos K, Cook I, Adams DJ, Taylor GR, Rabbitts P. 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 Res.* 2010 Aug;38(14):e151. doi: 10.1093/nar/gkq510. Epub 2010 Jun 4. PMID: 20525786; PMCID: PMC2919738.
- [Yao *et al.* (2023)] Yao H, Li H, Wang J, Wu T, Ning W, Diao K, Wu C, Wang G, Tao Z, Zhao X, Chen J, Sun X, Liu X-S. Copy number alteration features in pan-cancer homologous recombination deficiency prediction and biology. *Communications Biology.* 2023;6:527. doi:10.1038/s42003-023-04901-3
- [Yu *et al.* (2020)] Yu D, Liu Z, Su C, Han Y, Duan X, Zhang R, Liu X, Yang Y, Xu S. Copy number variation in plasma as a tool for lung cancer prediction using Extreme Gradient Boosting (XGBoost) classifier. *Thoracic Cancer.* 2020;11(1):95–102. doi:10.1111/1759-7714.13204
- [Zou and Hastie (2005)] Zou, H. and Hastie, T. (2005) Regularization and Variable Selection via the Elastic Net. *Journal of the Royal Statistical Society, Series B*, 67, 301-320. <https://doi.org/10.1111/j.1467-9868.2005.00503.x>