scientific reports

Check for updates

OPEN

Stochastic LASSO for extremely high-dimensional genomic data

Beomsu Baek[1], Jongkwon Jo[2,3], Mingon Kang[1]✉ & Youngsoon Kim[2]✉

Accurate identification of significant features in high-dimensional data is indispensable in high-throughput genomic analysis and association studies. Least Absolute Shrinkage and Selection Operator (LASSO) and its derivatives have been widely adapted to discover potential biomarkers as a feature selection scheme in various biological systems. Recently, bootstrap-based LASSO models, such as Random LASSO and Hi-LASSO, have been effective solutions for extremely high-dimensional but low sample size (EHDLSS) genomic data. However, the bootstrap-based LASSO models still have several drawbacks, such as multicollinearity within bootstrap samples, missing predictors in draw, and randomness in predictor sampling. To tackle the limitations, we propose a new bootstrap-based LASSO, named Stochastic LASSO, that effectively reduces multicollinearity in bootstrap samples and mitigates randomness in predictor sampling, resulting in remarkably outperforming benchmarks in feature selection and coefficient estimation. Furthermore, Stochastic LASSO provides a two-stage *t*-test strategy for selecting statistically significant features. The performance of Stochastic LASSO was assessed by comparing the existing benchmark models in extensive simulation experiments. In the simulation experiments, Stochastic LASSO consistently showed significant improvements in performance compared to the state-of-the-art LASSO models for feature selection, coefficient estimation, and robustness. We also applied Stochastic LASSO for the gene expression data of publicly available TCGA cancer datasets and identified statistically significant genes associated with survival month prediction. The source code is publicly available at: https://github.com/datax-lab/StochasticLASSO.

Keywords Stochastic LASSO, LASSO, High-dimensional data, Variable selection

Identification of a subset of significant predictors is indispensable in understanding biological mechanisms and enhancing predictive performance in high-throughput and high-dimensional genomic data[1–3]. Least Absolute Shrinkage and Selection Operator (LASSO)[4] and its derivatives have powered to identify a subset of relevant predictors in several biological applications, such as survival analysis[5], metabolomics[6], discovering protein–protein interactions[7], and Cox proportional hazard modeling[8]. However, conventional LASSO models, including Adaptive LASSO[9], Relaxed LASSO[10], and Precision LASSO[11], have following limitations when applied to extremely high-dimensional and low sample size (EHDLSS) data: (1) LASSO selects predictors only up to the sample size, and (2) LASSO does not identify all the predictors that are highly correlated with each other. These limitations primarily stem from the sparsity-inducing nature of the L_1 regularization. Although Elastic-Net[12] mitigates these issues by incorporating an L_2 regularization, the L_2 component introduces its own challenge in EHDLSS settings, as it forces highly correlated predictors that have opposite coefficient signs to be estimated with the same sign, thereby reducing feature interpretability. Thus, the conventional LASSOs are challenged to apply for omics datasets that include hundreds of patent samples of more than 20,000 genes or 80,000 SNPs with highly multicollinearity.

Bootstrap-based LASSOs, such as Random LASSO[13], Recursive Random LASSO[14], and Hi-LASSO[15], addressed the EHDLSS issues by drawing multiple bootstrap samples of lower-dimensionality and then aggregating the results for the final feature selection. Most bootstrap-based LASSOs consist of two procedures: (1) calculating importance scores of predictors for the oracle property[16] and (2) estimating coefficients of predictors by prioritizing predictors of high importance scores. The bootstrap-based LASSOs have the advantages of accurate feature selection, precise coefficient estimation, and enhanced predictive performance with EHDLSS data.

However, the bootstrap-based LASSOs also have several drawbacks, such as a multicollinearity issue within bootstrap samples, missing predictors in draw, and randomness in predictor sampling. First, multicollinearity

[1]Department of Computer Science, University of Nevada, Las Vegas 89154, NV, USA. [2]Department of Information and Statistics, Gyeongsang National University, Jinju, Republic of Korea. [3]AI Research Team, BigAI Inc., Changwon, Republic of Korea. ✉email: mingon.kang@unlv.edu; youngsoonkim@gnu.ac.kr

within bootstrap samples often causes underestimated importance scores of non-zero predictors. Although bootstrap-based LASSOs reduce multicollinearity in the entire set of predictors through bootstrap sampling, the bootstrap samples still have local multicollinearity. For instance, non-zero predictors are often estimated as zero, when highly correlated non-zero predictors are drawn in a bootstrap sample. Moreover, if highly correlated predictors are with opposite coefficient signs, only the predictors with dominant sign are identified. The predictors of a non-dominant sign are estimated as zero or incorrectly identified as the dominant sign. Second, bootstrap-based LASSOs have a potential risk of missing predictors during the bootstrap process. Let the model construct B numbers of bootstrap samples by drawing q predictors from a total of p predictors ($q < p$). The number of times that each predictor is drawn follows the binomial distribution. Then, when 500 bootstrap samples are generated by drawing 200 predictors from 20,000 predictors (e.g., $B = 500$, $q = 200$, and $p = 20,000$), the expected number of predictors that are never selected is 131.4, and the expected number of predictors selected less than three times is 2,467.7. Thus, a number of the predictors' coefficients would be missing or underestimated, since the majority of predictors are seldom drawn. Lastly, missing predictors or imbalanced predictor inclusion in the bootstrap, caused by randomness in predictor sampling, make coefficient estimations ineffective. Bootstrap-based LASSOs construct bootstrap samples by drawing q predictors with probabilities proportional to the importance scores. However, the randomness in draw still makes chances of several non-zero predictors missing or seldom consideration from the regression model in a bootstrap.

In this paper, we propose an enhanced bootstrap-based LASSO model, named Stochastic LASSO, which remarkably improves the current LASSO solutions. The main contributions of Stochastic LASSO are as follows:

- Stochastic LASSO significantly enhances the feature selection performance, as well as accurately estimating true coefficients, comparing to the state-of-the-art LASSO models,
- Stochastic LASSO proposes a parametric statistical test for selecting significant feature in high-dimensional data, and
- Stochastic LASSO produces robust feature selection results.

The rest of the paper is organized as follows: Section Methods describes the proposed Stochastic LASSO in detail, and in Section Experimental results, we conducted the assessment of Stochastic LASSO by comparing it with existing state-of-the-art LASSO models.

Methods

Overview

Stochastic LASSO follows the procedures: (1) constructing lower-dimensional, linearly independent bootstrap samples by a Correlation Based Bootstrapping (CBB) strategy, (2) estimating coefficients of each bootstrap sample, (3) calculating local scores for feature selection, (4) estimating the final coefficients (i.e., global scores) by forward selection, and (5) determining statistical significance by a two-stage t-test in a one-time bootstrapping procedure. Stochastic LASSO improves the LASSO solution by proposing (1) an enhanced bootstrapping algorithm for reducing multicollinearity (§2.2), (2) a forward selection strategy for robust feature selection and coefficient estimation (§2.3), and (3) a statistical strategy for identifying statistically significant features in high dimensional data (§2.4).

Reducing multicollinearity in bootstrap samples

Stochastic LASSO effectively reduces multicollinearity in bootstrap samples by the proposed Correlation Based Bootstrapping (CBB) algorithm. CBB penalizes predictors highly correlated with others in the bootstrapping so that the predictors of bootstrap samples become independent. CBB sets a selection probability of each predictor by the correlation with other predictors in the bootstrap sample, when Stochastic LASSO draws a predictor during bootstrap sampling. Let S and Q be the sets of the predictor indices, where S includes predictor indices that have not been drawn and Q is with indices already included in the bootstrap sample. Initially, $S = \{1, ..., p\}$ and $Q = \emptyset$. Then, CBB computes the selection probabilities of S based on correlation with Q. We refine the probability $Pr(p_i)$ (where $i \in S$) that the i-th predictor is selected, as follows:

$$Pr(p_i) = \frac{1}{\sum_{q \in Q} (r_{q,i})^2} \Big/ \sum_{s \in S} \frac{1}{\sum_{q \in Q} (r_{q,s})^2}, \tag{1}$$

where $r_{i,j}$ denotes the Pearson correlation coefficient between i-th and j-th predictors. Once an i-th predictor is randomly selected with $Pr(p_i)$, CBB updates S and Q as $S = S \backslash \{i\}$ and $Q = Q \cup i$ until the bootstrap sample is constructed with q predictors (i.e., $|Q| = q$), where $S \backslash \{i\}$ is the vector S excluding the element i. When Stochastic LASSO draws the first predictor of the bootstrap sample (i.e., $Q = \emptyset$), CBB sets $Pr(p_i)$ to $1/|S|$. Therefore, CBB prioritizes predictors that are likely independent to the other predictors in the bootstrapping.

In addition, Stochastic LASSO guarantees that all predictors are drawn equal times by sampling predictors without replacement. CBB algorithm constructs $\lceil p/q \rceil$ bootstrap samples until $S = \emptyset$, ensuring that each predictor is drawn once. To guarantee that each predictor is drawn exactly same times, Stochastic LASSO repeats the CBB algorithm r times. Consequently, Stochastic LASSO estimates r coefficients for each of the p predictors by applying penalized linear regression to each bootstrap samples. The CBB algorithm was implemented using Numpy's corrcoef() function in Python, which leverages an accelerated matrix multiplication routine based on the Basic Linear Algebra Subprograms (BLAS). We assessed the scalability of CBB with high-dimensional data for computational feasibility. The detailed complexity analysis of CBB is provided in Supplementary Note S1 and Table S1.

Improving coefficient estimation

Stochastic LASSO determines the optimal subset of features and estimates their coefficients using forward selection based on local scores. While traditional forward selection methods require evaluating all possible feature subsets, Stochastic LASSO defines a fixed subset based on the rank of local scores. Let $B = \{\hat{b}_{ij} \mid i = 1, \ldots, r,\ j = 1, \ldots, p\}$ be a $r \times p$ matrix that includes r coefficient estimates of p numbers of variables obtained from bootstrapping procedure. We define the local score of the j-th predictor (L_j) as follows:

$$L_j = \sum_{i=1}^{r} \left|\hat{b}_{ij}\right| / r, \tag{2}$$

Let $S_0 = \emptyset$ be an initial feature subset, and the subsequent subsets (S_k) are defined by prioritizing features with high local scores:

$$S_k = S_0 \cup \{x_1, \ldots, x_k\}, \tag{3}$$

where x_j denotes the predictor with the j-th highest local score. Then, Stochastic LASSO estimates the coefficients of the feature subsets (S_k) by applying penalized linear regression (e.g., Elastic-Net), and determines the optimal subset S^* as follows:

$$S^* = argmin_{S_k} SSE_{val}(S_k), \tag{4}$$

where $SSE_{val}(S_k)$ denotes the sum of squared error of validation data calculated by the estimated coefficients of S_k. Thus, Stochastic LASSO systematically constructs a subset consisting of non-zero predictors and precisely estimates their partial correlations, without additional bootstrapping procedures that most bootstrap-based LASSOs require.

Tests of significance for statistically significant feature

Stochastic LASSO proposes a statistical test, two-stage t-test (TSTT), for evaluating statistical significance of features. TSTT sequentially conducts one-sample t-test and two-sample t-test. The one-sample t-test identifies potential significant features whose coefficient means are non-zeros, and the two-sample t-test selects significant features with relatively large coefficient magnitudes among the potential significant features. Let β_j be a regression coefficient of the j-th feature. The coefficient estimate of the j-th feature, $\hat{\beta}_j$, is defined as $\sum_{i=1}^{r} \hat{\beta}_{ij}/r$, where $\hat{\beta}_{ij}$ denotes the i-th coefficient estimate of the j-th feature. The one-sample t-test identifies the potential significant features under the following hypotheses:

$$H_0 : \beta_j = 0,\ H_1 : \beta_j \neq 0. \tag{5}$$

Let K be the set of indices of the m potential significant features selected by the previous one-sample t-test. The coefficient magnitudes of the j-th feature ($j \in K$) and all potential significant features (i.e.,K), denoted as $|\beta_j|$ and $|\beta_K|$, are estimated as $\sum_{i=1}^{r} \left|\hat{\beta}_{ij}\right| / r$ and $\sum_{j \in K} \sum_{i=1}^{r} \frac{\left|\hat{\beta}_{ij}\right|}{m \times r}$, respectively. Then significant features were determined using the two-sample t-test under the following hypotheses:

$$H_0 : |\beta_j| \leq |\beta_K|,\ H_1 : |\beta_j| > |\beta_K|, \tag{6}$$

The details of TSTT including test statistics and rejection regions are elucidated in Supplementary Note S2.

The proposed TSTT allows Stochastic LASSO to identify more features than the sample size, by individually assessing the statistical significance of each feature based on the bootstrap-derived coefficient estimates. Therefore, Stochastic LASSO not only reliably evaluates the statistical significance of features but also fundamentally addresses the inherent limitations of conventional LASSO models by applying TSTT in extremely high-dimensional settings. The detail procedures of Stochastic LASSO are described in Algorithm 1.

1. Draw bootstrap samples, consisting of n samples of q predictors until each predictor is selected exactly r times by the Correlation Based Bootstrapping (CBB) algorithm.
2. Estimate coefficients of predictors, $\{\hat{b}_{ij} \,|\, i = 1, \ldots, r,\ j = 1, \ldots, p\}$, by applying penalized linear regression (e.g., Elastic-Net) to each bootstrap sample after z-score normalization.
3. Compute local scores by $L_j = \sum_{i=1}^{r} \left|\hat{b}_{ij}\right| / r$.
4. Construct subsets of predictors, $S_k = \{x_1, \ldots, x_k\}$, by prioritizing predictors with high local scores, where x_j denotes the predictor with the j-th highest local score.
5. Determine the optimal subset, S^*, that minimizes the prediction error by estimating the final coefficients of S_k through penalized linear regression (e.g., Elastic-Net).
6. Select statistically significant predictors with the significance level α (e.g., 0.05 or 0.01), using a two-stage t-test.

Algorithm 1. Stochastic LASSO.

Experimental results

We assessed the performance of Stochastic LASSO with the various experimental settings, compared to state-of-the-art LASSO models on the following criteria: (1) feature selection, (2) coefficient estimation, (3) test of significance, and (4) model robustness. First, we evaluated how accurately Stochastic LASSO identifies non-zero variables with various scales of data dimensionalities. Second, we assessed the performance of coefficient estimation by computing the residual sum between the estimations and ground truth. Third, we examined whether Stochastic LASSO can identify statistical significance of non-zero variables. Lastly, we evaluated the consistency and robustness of feature selection.

Feature selection

For the assessment of feature selection, we conducted simulation studies where ground truth of non-zero variables are known. The simulation study mainly considered high-dimensional, but low-sample-size data with high multicollinearity, which is a common setting in genomic data. We followed the simulation settings that have been commonly used in most bootstrap-based LASSO studies[13–15]. We generated four synthetic datasets (i.e., Dataset I-IV) with the following linear regression model:

$$y = \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p + \epsilon, \tag{7}$$

where $\epsilon \sim N(0, \sigma^2)$, $x_i \sim N(0, 1)$, and various types of multicollinearity were introduced by predefined covariance matrices. The four dataset consisted of varying numbers of variables and samples, defined in Table 1. Dataset I included 50 samples of 100 variables, where the first ten coefficients were set to non-zeros. The regression coefficients of ground truth were defined as:

$$\beta = (3, 3, -3, 2, 2, -2, 1.5, 1.5, 1.5, -1.5)\,, \tag{8}$$

and the multicollinearity was introduced by predefined covariance matrix:

	Sample sizes		Numbers of features		
	Training	Validation*	Total	Non-zeros	β
Dataset I	50	10	100	10	$(3, 3, -3, 2, 2, -2, 1.5, 1.5, 1.5, -1.5, 0, \ldots, 0)$
Dataset II	100	20	1,000	50	$\beta_1, \ldots, \beta_{50} \sim N(0, 4)$, others zero
Dataset III	200	40	10,000	50	$\beta_1, \ldots, \beta_{50} \sim N(0, 4)$, others zero
Dataset IV	400	80	10,000	50	$\beta_1, \ldots, \beta_{50} \sim N(0, 4)$, others zero

Table 1. Description of the simulation data. *The validation data is further generated with the size equivalent to 20% of the training data.

$$\begin{bmatrix} \Sigma_{0.9}^{3} & 0 & 0 & 0 \\ 0 & \Sigma_{0.9}^{3} & 0 & 0 \\ 0 & 0 & \Sigma_{0.9}^{4} & 0 \\ 0 & 0 & 0 & I^{90} \end{bmatrix}, \tag{9}$$

where Σ_v^k is a $k \times k$ matrix with unit diagonal elements and off-diagonal elements of value v, and I^k is an identity matrix of size k. Dataset II consists of 100 samples of 1,000 variables, where the first 50 non-zero coefficients were drawn from $N(0, 4)$. The multicollinearity with high and low degrees was designed using the following covariance matrix:

$$\begin{bmatrix} \Sigma_{0.9}^{15} & 0 & 0 & 0 \\ 0 & \Sigma_{0.9}^{15} & J_{0.3} & 0 \\ 0 & J_{0.3}^{T} & \Sigma_{0.9}^{20} & 0 \\ 0 & 0 & 0 & I^{950} \end{bmatrix}, \tag{10}$$

where J_v is a matrix with all unit elements of a value v. Dataset III includes 200 samples of 10,000 variables, where the first 50 coefficients of non-zero were drawn from $N(0, 4)$. The covariance matrix of Dataset III is as follows:

$$\begin{bmatrix} \Sigma_{0.9}^{15} & 0 & 0 & 0 \\ 0 & \Sigma_{0.9}^{15} & J_{0.3} & 0 \\ 0 & J_{0.3}^{T} & \Sigma_{0.9}^{20} & 0 \\ 0 & 0 & 0 & I^{9950} \end{bmatrix}, \tag{11}$$

In Dataset IV, we considered double samples in the same model of Dataset III. To tune the hyper-parameters of the LASSO models, we further generated validation data for each dataset, with a size equivalent to 20% of the training samples. Note that test data was not considered, as we evaluated only feature selection performance without assessing predictive errors.

The benchmark models include LASSO[4], Elastic-Net[12], Adaptive LASSO[9], Relaxed LASSO[10], Precision LASSO[11], Random LASSO[13], Recursive Random LASSO[14], and Hi-LASSO[15]. For the non-bootstrap-based LASSO models (i.e., LASSO, Elastic-Net, Adaptive LASSO, Relaxed LASSO, and Precision LASSO), hyper-parameters (e.g., regularization parameter) were tuned by minimizing prediction error on the validation data. For the bootstrap-based LASSO models of Random LASSO, Recursive Random LASSO, and Hi-LASSO, the number of variables (q_1 and q_2) were set to the sample number (i.e., n), and the number of bootstrap samples (B) was set to $p/q \times 30$ to ensure that each variable is included in the bootstrap samples 30 times on average. In Stochastic LASSO, we also set $q = n$ and $r = 30$ for the fair comparison. This configuration was applied as the default setting for all bootstrap-based LASSO models, and its rationale is supported by the sensitivity analysis of the Stochastic LASSO hyperparameters presented in Supplementary Note S3.

We computed F1-scores and Area Under the Precision-Recall Curve (AUCPR) for the evaluation. We defined non-zero variables as positive, and zero variables as negative. Then, the confusion matrices were computed as follows: True Positive (*TP*) if a model correctly identifies non-zero variables as non-zeros; False Positive (*FP*) if a model incorrectly identifies zero variables as non-zeros; False Negative (*FN*) if a model incorrectly identifies non-zero variables as zeros; and True Negative (*TN*) if a model correctly identifies zero variables as zeros. F1-score was calculated by $2(Precision \times Recall)/(Precision + Recall)$, where *Precision* and *Recall* are defined as $TP/(TP + FP)$ and $TP/(TP + FN)$, respectively. The AUCPR was calculated from the Precision-Recall curve, which is generated by evaluating the model across varying thresholds. We repeated the experiments ten times on randomly generated synthetic data for the reproducibility of the model performance.

Stochastic LASSO outperformed all other benchmarks across the synthetic datasets (Fig. 1A, Supplementary Table S2), showing the highest F1-scores of 0.7093±0.0197, 0.6650±0.0139, 0.5251±0.0170, and 0.7777±0.0046 with Dataset I-IV, respectively, which showed 8%, 77%, 65%, and 42% improvements against the second-best models. The outperformance of Stochastic LASSO to the second-best benchmark was statistically validated by the Wilcoxon rank-sum test (p-values<0.05) with all synthetic datasets (Supplementary Table S2). Furthermore, we verified the feature selection performance using AUCPR without thresholding. Stochastic LASSO also achieved the highest AUCPR of 0.8284±0.0095, 0.6992±0.0021, 0.5772±0.0089, and 0.6985± 0.0018 with Dataset I-IV, respectively, representing statistically significant (p-values<0.05) improvements of 15%, 37%, 31%, and 8% over the second-best models (Fig. 1B, Supplementary Table S2). We observed the remarkable performance of bootstrap-based LASSO models, including Random LASSO and Hi-LASSO, in the high dimensional data (e.g., Dataset II-IV), which implies that bootstrap-based LASSO is more suitable for extremely high-dimensional data analysis than non-bootstrap-based LASSO. However, Recursive Random LASSO showed the lowest F1-scores, despite being a bootstrap-based model, due to introducing initial biases when handling high-dimensional data.

Coefficient estimation

In this experiment, we verified how precisely the models estimate the ground truth coefficient values. The performance of coefficient estimation was evaluated by computing Root Mean Squared Errors (RMSE) between estimated coefficients and the ground truth in the synthetic datasets that we used in the previous experiment (Dataset I-IV). We computed $RMSE_{ALL}$ and $RMSE_{Nonzeros}$, where $RMSE_{ALL}$ was computed with all coefficients, and $RMSE_{Nonzeros}$ was computed with only non-zero coefficients. Stochastic LASSO showed the least errors to estimate coefficient values among the benchmark models (Fig. 2, Supplementary Table S3),

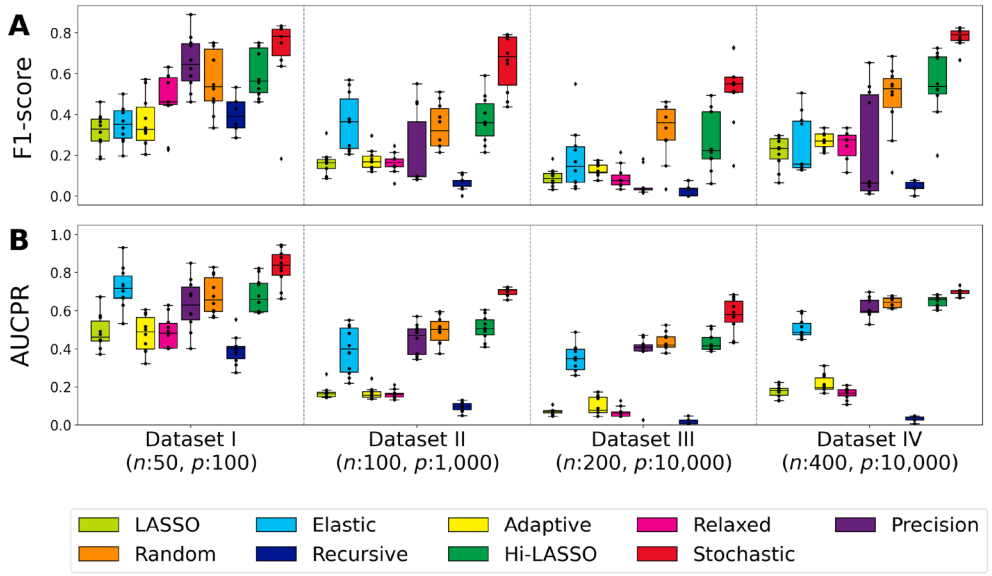

Fig. 1. Feature selection performance with simulation data. (**A**) Comparison of F1-score, (**B**) Comparison of AUCPR. On the x-axis of the figure, we labeled the sample size (n) and dimension (p) for each dataset.

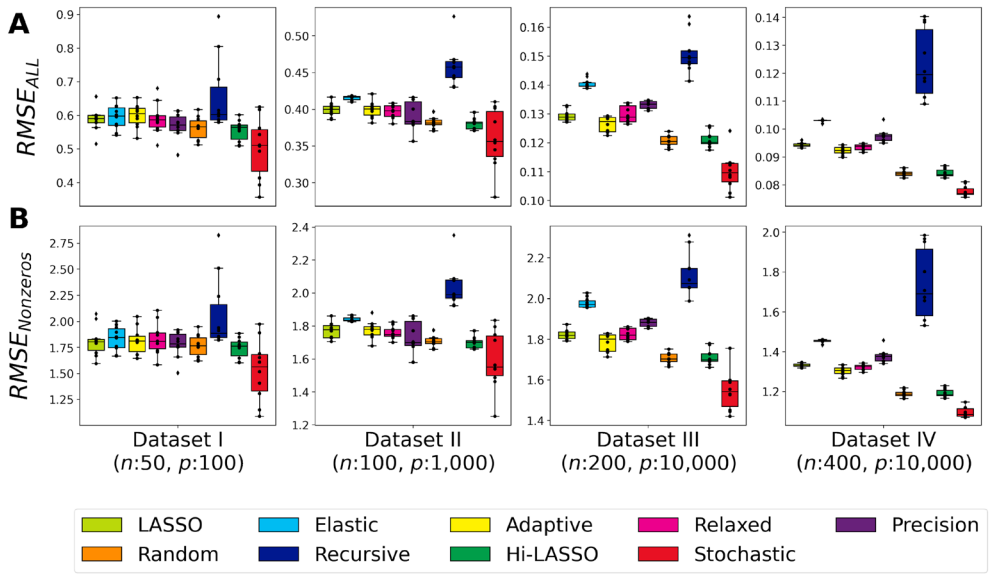

Fig. 2. Coefficient estimation performance with simulation data. (**A**) Comparison of $RMSE_{ALL}$, (**B**) Comparison of $RMSE_{Nonzeros}$. On the x-axis of the figure, we labeled the sample size (n) and dimension (p) for each dataset.

achieving the lowest $RMSE_{ALL}$ of 0.5022±0.0091, 0.3596±0.0041, 0.1099±0.0006, and 0.0777±0.0002 with Dataset I-IV, respectively, which represented 9%, 5%, 9% and 8% improvements over the second-best models. Stochastic LASSO also exhibited the lowest $RMSE_{Nonzeros}$ of 1.5286±0.0294, 1.5931±0.0179, 1.5464±0.0098, and 1.0945±0.0025 with Dataset I-IV, respectively, showing 12%, 6%, 9%, and 8% improvements against the second-best models. Stochastic LASSO's improvements for coefficient estimation were statistically validated compared to the second-best models (p-values<0.05). The second-best models were Hi-LASSO in Dataset I&II and Random LASSO in Dataset III&IV.

Furthermore, we explored the signs of Stochastic LASSO's coefficients compared to the second-best models of Hi-LASSO and Random LASSO. Fig. 3 depicts the results of the coefficient estimations of the models. In Fig. 3, the circle marker in black presents the ground truth of non-zero coefficients, while colored shape markers indicate the estimations of the models on average (i.e., Stochastic LASSO with pentagon in red, Hi-LASSO with square in green, and Random LASSO with triangle in orange). The figure shows that Stochastic LASSO more accurately estimated the signs of the coefficients than the others. In Dataset I, Random LASSO and Hi-LASSO

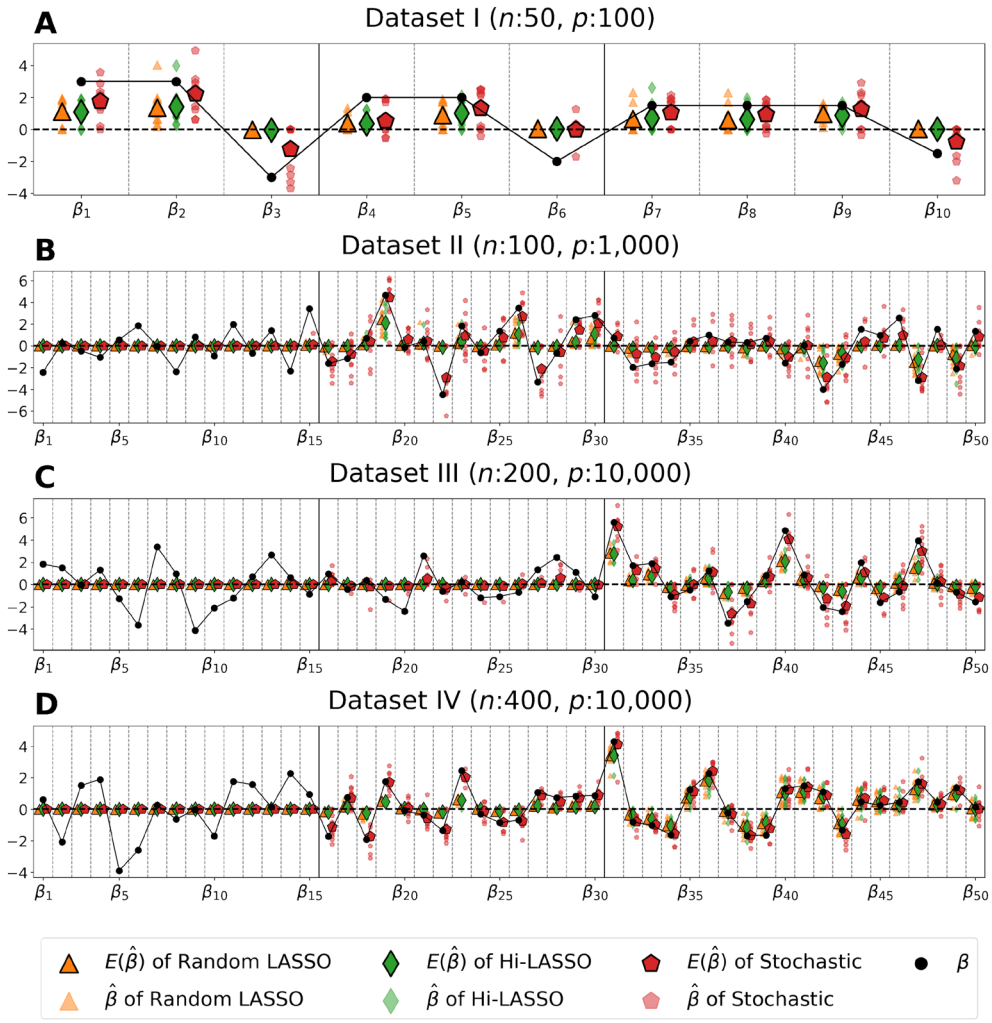

Fig. 3. Coefficient estimation with simulation data.

failed to estimate the negative signs on β_3, β_6, and β_{10}, whereas Stochastic LASSO successfully estimated them (Fig. 3A). Note that the non-zero variables in Dataset I are designed with predominantly positive coefficients and high multicollinearity (Eqs. 8 & 9). Consequently, the benchmark LASSO models barely identified the negative coefficients (i.e., β_3, β_6, and β_{10}) in 10 repeated experiments: Random LASSO estimated β_3, β_6, and β_{10} as negative coefficients 1, 0, and 0 times, respectively, and Hi-LASSO estimated them 2, 0, and 0 times, respectively. However, Stochastic LASSO demonstrated its notable coefficient estimation capability by accurately estimating β_3, β_6, and β_{10} as negative coefficients 4, 2, and 5 times, respectively. We also observed that Random LASSO and Hi-LASSO tend to estimate the same signs in highly collinearity. For instance, In Dataset II, Random LASSO and Hi-LASSO identified only the positive coefficients on $\beta_{16} - \beta_{30}$, and only the negative coefficients on $\beta_{31} - \beta_{50}$, while Stochastic LASSO correctly estimated the signs (Fig. 3B). In Datasets III-IV, only Stochastic LASSO precisely identified $\beta_{16} - \beta_{30}$, while Random LASSO and Hi-LASSO underestimated $\beta_{16} - \beta_{30}$ close to zero (Fig. 3C-D).

Tests of significance

We performed a semi-simulation study to assess whether the proposed tests of significance can identify the statistical significance of non-zero variables in EHDLSS data. In this study, we generated semi-simulation data, adapting gene expressions of 18 types of cancer in the TCGA databases. The gene expression data (i.e., RNA-seq) were directly used as the independent variables, but the dependent variable (e.g., survival month) was synthetically generated as follows: (1) We conducted a correlation analysis between gene expression and the survival months; (2) We selected 100 genes with the highest Pearson correlation coefficient with survival months; (3) The regression coefficients (β) of the 100 genes were randomly generated from the normal distribution, $N(0, 4)$; (4) The coefficients of the other genes were set to 0; (5) The dependent variables were generated from the linear combination of the gene expression (X), the coefficients (β), and the errors (ϵ) from the normal distribution with a mean of zero and the standard deviation of the logarithmic survival months (i.e., $y = X + \epsilon$). The semi-synthetic cancer datasets are briefly summarized in Table 2. In this experiment, we considered only Recursive Random LASSO and Hi-LASSO, which proposed tests of significance scheme for feature selection, as

Cancer type	Description	*n*	*p*
BRCA	Breast Invasive Carcinoma	1,082	20,192
COAD	Colorectal Adenocarcinoma	588	17,496
UCEC	Uterine Corpus Endometrial Carcinoma	526	17,496
LGG	Brain Lower Grade Glioma	513	20,145
KIRC	Kidney Renal Clear Cell Carcinoma	510	20,193
LUAD	Lung Adenocarcinoma	501	20,090
THCA	Thyroid Carcinoma	498	20,031
PRAD	Prostate Adenocarcinoma	493	19,072
LUSC	Lung Squamous Cell Carcinoma	478	20,167
SKCM	Skin Cutaneous Melanoma	428	20,160
STAD	Stomach Adenocarcinoma	407	16,755
BLCA	Bladder Urothelial Carcinoma	406	20,164
LIHC	Liver Hepatocellular Carcinoma	365	20,018
OV	Ovarian Serous Cystadenocarcinoma	299	19,045
CESC	Cervical Squamous Cell Carcinoma	294	20,002
KIRP	Kidney Renal Papillary Cell Carcinoma	282	20,103
LAML	Acute Myeloid Leukemia	161	16,754
GBM	Glioblastoma Multiforme	159	19,787

Table 2. Description of the semi-simulation data.

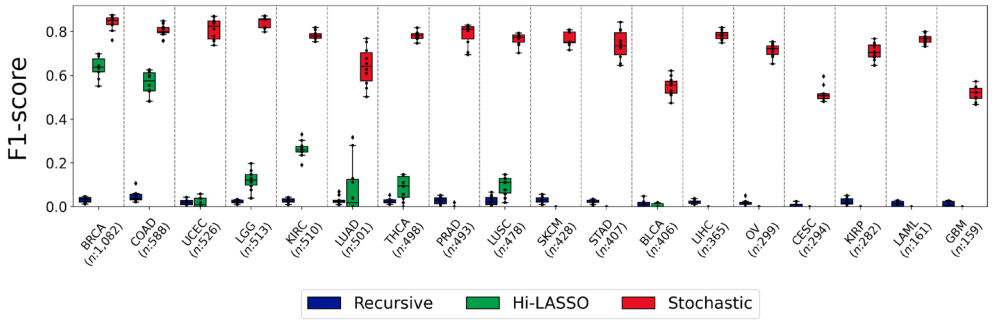

Fig. 4. F1-score of the semi-simulation study. The cancer type and sample size (n) for each dataset are described on the x-axis of the figure. The number of features (p) of each dataset are around 20,000.

a benchmark. In Stochastic LASSO, the hyperparameter q was set to the sample size of each cancer type, and r was set to 30 to ensure the normality of coefficient estimates. For a fair comparison, the hyperparameters q_1 and q_2 were set to the sample size, and B was set to $p/q \times 30$ in Recursive Random LASSO and Hi-LASSO. Although the tests of significance were applied to identify statistically significant variables rather than to estimate their coefficients, the experimental results with RMSE of coefficient estimations are also provided in Supplementary Tables S4–S5. Since a p-value indicates the statistical significance of individual predictors rather than a ranking among predictors, the feature selection performance was assessed solely based on the F1-score at a significance level of $\alpha = 0.05$.

Stochastic LASSO precisely identified statistically significant non-zero variables, achieving the highest F1-scores across the 18 datasets (Fig. 4, Supplementary Table S6). Stochastic LASSO produced F1-scores of above 0.8 for Breast Invasive Carcinoma (BRCA), Colorectal Adenocarcinoma (COAD), Uterine Corpus Endometrial Carcinoma (UCEC), and Brain Lower Grade Glioma (LGG) datasets, each of which consists of more than 500 samples. Stochastic LASSO also maintained F1-scores of over 0.5 for Glioblastoma Multiforme (GBM) datasets, where the number of samples (n : 159) is extremely small compared to the number of features (p : 19,787). In contrast, the second-best Hi-LASSO's performance distinctly declined as the sample sizes decreased. Hi-LASSO showed F1-scores of over 0.5 in the largest sample size datasets only (i.e, BRCA and COAD). These experimental results demonstrate that Stochastic LASSO can provide a reliable feature selection even in extremely high dimensional settings.

Furthermore, we conducted an additional experiment using the GBM dataset, which had the smallest sample size among the semi-simulation data, to evaluate whether Stochastic LASSO can select more variables than the sample size. The overall experimental settings were consistent with the previous semi-simulation study, but 500 genes were assigned non-zero coefficients, which exceeded the sample size ($n = 159$). As a result, Stochastic LASSO selected an average of 307.6 variables across 10 repetitions (Fig. 5A), demonstrating that it overcomes a

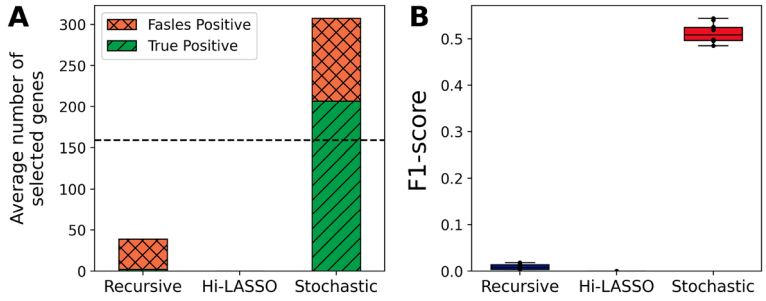

Fig. 5. Feature selection results on the TCGA GBM dataset. (**A**) Average number of selected genes across 10 repetitions. The dashed line indicates the sample size of the GBM dataset ($n = 159$). (**B**) Comparison of F1-score.

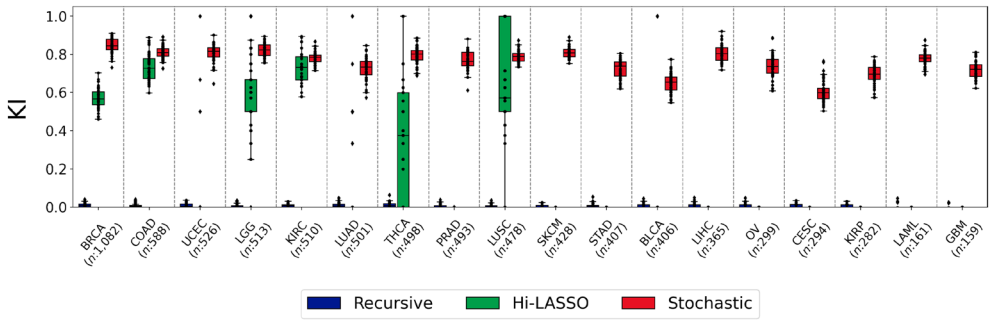

Fig. 6. Kuncheva Index (KI) of the semi-simulation study. The cancer type and sample size (n) for each dataset are described on the x-axis of the figure. The number of features (p) of each dataset is around 20,000.

key limitation of conventional LASSO models. Specifically, Stochastic LASSO identified 206.8 non-zero (True Positive) and 100.8 zero variables (False Positive) on average, while Recursive Random LASSO selected 2.2 non-zero and 36.7 zero variables, and Hi-LASSO did not detect any significant genes. Furthermore, Stochastic LASSO achieved the highest F1-score of 0.5120±0.006 (Fig. 5B), which is consistent with the previous feature selection results on the GBM dataset, indicating that the proposed two-stage t-test (TSTT) performs reliably across various high-dimensional settings.

Robustness of feature selection

We finally assessed whether our bootstrap-based LASSO can produce consistent feature selection performance by computing a pair-wise Kuncheva Index (KI) on the previous semi-synthetic cancer datasets. KI computes a robustness score in the range of [-1, 1], where zero indicates that each selection was made independently, a positive value indicates the feature selection model produces a stable selection, and a negative value indicates the model is unstable[17]. The equation of KI is:

$$KI(s_i, s_j) = \frac{|s_i \cap s_j| - \frac{|s_i| \times |s_j|}{p}}{min\left(|s_i|, |s_j|\right) - max\left(0, |s_i| + |s_j| - p\right)}, \tag{12}$$

where s_i and s_j are two sets of feature selected by the model, and p is the dimensionality of the dataset.

Stochastic LASSO produced the most consistent feature selection with the highest average KI of 0.7590 with smallest variances across 18 cancer types, whereas Hi-LASSO and Recursive Random LASSO showed 0.2158 and 0.0058, respectively (Fig. 6, Supplementary Table S7). Note that Stochastic LASSO achieved the highest F1-scores on the same datasets. The highest F1-scores and KI demonstrate Stochastic LASSO's reliable feature selection capability for extremely high-dimensional genomic data.

Glioblastoma & Glioma gene expression data analysis

We applied Stochastic LASSO for Glioblastoma Multiforme (GBM) & Brain Lower Grade Glioma (LGG) to assess it with real-world biological data, where ground truths are not known. GBM is the most aggressive and common primary brain tumor in adults, while LGG is a slower-growing, less aggressive brain tumor type with a generally better prognosis[18]. We downloaded genomic data of GBM and LGG from The Cancer Genome Atlas Program (TCGA) and combined them. In the experiment, we used the gene expressions (i.e., RNA-seq) of GBM & LGG patients as independent variables and their survival time as the dependent variable. The gene expression data consist of 266 patients and 19,785 genes, including only deceased patients for the regression problem rather

Gene	Coefficient	Correlation	Reference
ONECUT3	3.7334	0.4760	[19]
HILS1	-3.6233	-0.2751	[20–22]
GSC2	3.4914	0.4509	-
AMTN	3.4667	0.4520	-
KRTAP24-1	-3.4309	-0.4247	-
TEX37	3.4101	0.4272	-
DEFB120	3.3903	0.5019	-
UCN3	3.3361	0.4519	[23,24]
BHLHE23	3.1976	0.4389	-
LINC00442	3.1052	0.4495	-
SNAR-F	3.0399	0.3547	-
COL22A1	-3.0373	-0.3024	[25–27]
LINC00114	-3.0349	-0.4191	[28]
OR51I2	-3.0318	-0.4248	-
OR51Q1	3.0154	0.3538	-
UGT1A10	2.9891	0.3797	-
VN1R4	2.9794	0.4220	[29]
LOC642929	2.9704	0.4202	-
HELT	2.9660	0.3857	-
TEX13A	2.9124	0.4181	-

Table 3. Top-20 ranked genes by Stochastic LASSO in GBM & LGG. Note: The 20 top-ranked genes, sorted in descending order of the absolute value of their estimated coefficients, along with the Pearson correlations with survival time, are presented.

than cox-regression. The average survival time was 28.1 months. We identified genes related to the survival time using Stochastic LASSO, where q was set to 266 (i.e., number of sample) and r was set to 30 to ensure the normality of coefficient estimates. For feature selection, a significance level of 0.05 was used for the statistical tests in Stochastic LASSO.

Stochastic LASSO identified 490 statistically significant genes out of 19,785 genes, and we examined the genes in the biological literature. Consequently, a number of genes are shown as associated with biomarkers in GBM & LGG (Table 3). For instance, ONECUT3 was reported to suppress glioblastoma cell proliferation and promote a glial-to-neuronal identity switch, implicating it in GBM reprogramming[19]. HILS1 was identified as a strong prognostic biomarker for LGG. It was significantly upregulated in glioma tissues and associated with higher tumor grade and worse survival outcomes[20,21]. HILS1 was also a part of the five-pseudogene prognostic signatures for lower-grade gliomas, with higher expression levels associated with advanced tumor grade and poorer patient survival[22]. Our analysis showed a negative coefficient, aligning with the report. UCN3 was found to be upregulated upon serum stimulation in glioma cells and transiently increased following proliferative stimuli, suggesting a role in glioma adaptation to environmental stress[23,24]. COL22A1 was overexpressed in GBM and identified as a key angiogenesis-related gene in grade 4 diffuse gliomas. Elevated COL22A1 expression was consistently associated with poor overall survival and endothelial remodeling in the GBM microenvironment[25–27]. LINC00114 was implicated in temozolomide resistance in GBM through ceRNA network regulation[28]. VN1R4 exhibited recurrent genomic alterations in glioma, particularly in LGG, suggesting its association with gliomagenesis and structural genome instability[29].

Furthermore, we conducted gene set enrichment analysis (GSEA) using 490 genes identified by Stochastic LASSO, based on 844 curated pathways retrieved from the Kyoto Encyclopedia of Genes and Genomes (KEGG) database[30]. The statistical significance of enrichment was assessed using the hypergeometric test. Consequently, 13 significantly enriched pathways (p-value<0.05) were identified, of which seven pathways have been previously reported to be closely associated with the prognosis of GBM and LGG (Table 4). The olfactory transduction pathway was identified as the most prominently enriched in long-term glioblastoma survivors, as revealed by transcriptomic and epigenetic analyses[31]. Translation initiation was found to be overrepresented in glioblastoma, indicating its relevance to tumor progression and patient outcomes[32]. The ribosome pathway was reported to be significantly downregulated in glioblastoma cells with acquired temozolomide resistance[33]. Retinol metabolism was also downregulated in glioma, suggesting that dysregulation in retinoid processing may contribute to remodeling of the tumor immune microenvironment[34]. The drug metabolism – cytochrome P450 pathway showed notable enrichment in glioma-associated hub genes, indicating a potential link to altered drug response and resistance mechanisms[35]. The metabolism of xenobiotics by cytochrome P450 pathway was enriched in glioblastoma, implicating detoxification processes in therapeutic resistance[36]. The JAK/STAT signaling pathway was reported to play a central role in glioma pathobiology by regulating tumor growth, invasion, immune evasion, and stemness, and thus represented a potential therapeutic target[37].

Pathway	Pathway genes	Selected genes	P-values	Reference
Olfactory Transduction	389	44	1.11e-16	[31]
Translation Initiation	81	13	9.11e-08	[32]
Ribosome	88	13	2.48e-07	[33]
Ascorbate and Aldarate Metabolism	25	5	3.22e-04	-
Pentose and Glucuronate Interconversions	28	5	5.60e-04	-
Retinol Metabolism	64	7	1.01e-03	[34]
Drug Metabolism other Enzymes	51	6	1.56e-03	-
Drug Metabolism-Cytochrome P450	72	7	2.01e-03	[35]
Steroid Hormone Biosynthesis	55	6	2.32e-03	-
Porphyrin and Chlorophyll Metabolism	41	5	3.28e-03	-
Metabolism of Xenobiotics by Cytochrome P450	70	6	7.71e-03	[36]
Starch and Sucrose Metabolism	52	5	9.12e-03	-
JAK/STAT Signaling pathway	155	8	3.94e-02	[37]

Table 4. Enriched pathway by Stochastic LASSO in GBM & LGG. Note: Significantly enriched pathways identified by Stochastic LASSO (p-value<0.05) are listed, along with the total number of pathway-related genes and the number of genes selected by Stochastic LASSO.

Conclusions

In this study, we have proposed Stochastic LASSO, an enhanced LASSO model for feature selection with high-throughput data. Stochastic LASSO improves the bootstrap-based LASSO models by: (1) reducing multicollinearity within bootstrap samples while ensuring that each predictor is included an equal number of times, (2) mitigating randomness in predictor sampling by forward selection without additional bootstrapping procedures, and (3) improving statistical significance tests with the proposed two-stage *t*-test. The performance of Stochastic LASSO was compared to the state-of-the-art LASSO models in extensive simulation settings and with real genomic data. Stochastic LASSO outperformed the benchmarks for feature selection, coefficient estimation, tests of significance, and robustness in the experiments. Stochastic LASSO was applied to gene expression data from TCGA GBM and LGG, identifying both statistically significant genes and enriched pathways associated with survival outcome. As a general framework applicable to any linear regression–based model, Stochastic LASSO can be extended to survival analysis models such as the Cox proportional hazards model, as well as to other applications including protein–protein interaction analysis and association studies. Furthermore, Stochastic LASSO can be applied to classification analysis using penalized logistic regression.

Data availability

The datasets analyzed during the current study are all publicly available online from The Cancer Genome Atlas (TCGA). The open-source is available at: https://github.com/datax-lab/StochasticLASSO.

Received: 23 June 2025; Accepted: 5 January 2026
Published online: 14 January 2026

References

1. Wu, T. T., Chen, Y. F., Hastie, T., Sobel, E. & Lange, K. Genome-wide association analysis by lasso penalized logistic regression. *Bioinformatics* **25**(6), 714–721 (2009).
2. Xu, C., Fang, J., Shen, H., Wang, Y. P. & Deng, H. W. EPS-LASSO: Test for high-dimensional regression under extreme phenotype sampling of continuous traits. *Bioinformatics* **34**(12), 1996–2003 (2018).
3. Geeven, G., van Kesteren, R. E., Smit, A. B. & de Gunst, M. C. M. Identification of context-specific gene regulatory networks with GEMULA-gene expression modeling using LAsso. *Bioinformatics* **28**(2), 214–221 (2012).
4. Tibshirani, R. Regression Shrinkage and Selection Via the Lasso. *J. R. Stat. Soc., B: Stat. Methodol.* **58**(1), 267–288 (1996).
5. Wang, W. & Liu, W. Integration of gene interaction information into a reweighted Lasso-Cox model for accurate survival prediction. *Bioinformatics* **36**(22–23), 5405–5414 (2020).
6. Fu, G. H., Yi, L. Z. & Pan, J. LASSO-based false-positive selection for class-imbalanced data in metabolomics. *J. Chemom.* **33**(10), e3177 (2019).
7. Yu, B. et al. Prediction of protein-protein interactions based on elastic net and deep forest. *Expert Syst. Appl.* **176**, 114876 (2021).
8. Sohn, I., Kim, J., Jung, S. H. & Park, C. Gradient lasso for Cox proportional hazards model. *Bioinformatics* **25**(14), 1775–1781 (2009).
9. Zou, H. The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.* **101**(476), 1418–1429 (2006).
10. Meinshausen, N. Relaxed Lasso. *Comput. Stat. Data Anal.* **52**(1), 374–393 (2007).
11. Wang, H., Lengerich, B. J., Aragam, B. & Xing, E. P. Precision Lasso: Accounting for correlations and linear dependencies in high-dimensional genomic data. *Bioinformatics* **35**(7), 1181–1187 (2019).
12. Zou, H. & Hastie, T. Regularization and variable selection via the elastic net. *J. R. Stat. Soc., B: Stat. Methodol.* **67**(2), 301–320 (2005).
13. Wang, S., Nan, B., Rosset, S. & Zhu, J Random lasso. *Ann. Appl. Stat.* **5**(1), 468–485, (2011).
14. Park, H., Imoto, S. & Miyano, S. Recursive random lasso (RRLasso) for identifying anti-cancer drug targets. *PLoS ONE* **10**(11), e0141869 (2015).
15. Kim, Y., Hao, J., Mallavarapu, T., Park, J. & Kang, M. Hi-LASSO: High-Dimensional LASSO. *IEEE Access* **7**, 44562–44573 (2019).

16. Fan, J. & Li, R. Variable selection via nonconcave penalized likelihood and its oracle properties. *J. Am. Stat. Assoc.* **96**(456), 1348–1360 (2001).
17. Lustgarten, J. L., Gopalakrishnan, V. & Visweswaran, S. Measuring stability of feature selection in biomedical datasets. *AMIA Annu. Symp. proc.* **406–410**, 2009 (2009).
18. Das, N. D. et al. Defining super-enhancers by highly ranked histone H4 multi-acetylation levels identifies transcription factors associated with glioblastoma stem-like properties. *BMC Genomics* **24**(1), 574 (2023).
19. Zupančič, M. et al. Concerted transcriptional regulation of the morphogenesis of hypothalamic neurons by ONECUT3. *Nat. Commun.* **15**(1), 8631 (2024).
20. Lee, M. An ensemble deep learning model with a gene attention mechanism for estimating the prognosis of low-grade glioma. *Biology (Basel)* **11**(4), 586 (2022).
21. Wang, Y. et al. Identification of a five-pseudogene signature for predicting survival and its ceRNA network in glioma. *Front. Oncol.* **9**, 1059 (2019).
22. Liu, B. et al. A prognostic signature of five pseudogenes for predicting lower-grade gliomas. *Biomed. Pharmacother.* **117**, 109116 (2019).
23. Akiyoshi, K. et al. Expression of mRNAs of Urocortin in the STKM-1 gastric cancer cell line. *Anticancer Res.* **33**(12), 5289–5294 (2013).
24. Kamada, M. et al. Expression of mRNAs of urocortin and corticotropin-releasing factor receptors in malignant glioma cell lines. *Anticancer Res.* **32**(12), 5299–5307 (2012).
25. Yan, B. et al. Artificial intelligence-based radiogenomics reveals the potential immunoregulatory role of COL22A1 in glioma and its induced autoimmune encephalitis. *Front. Immunol.* **16**, 1562070 (2025).
26. Liu, H., Zeng, Z. & Sun, P. Prognosis and immunoinfiltration analysis of angiogene-related genes in grade 4 diffuse gliomas. *Aging (Albany NY)* **15**(18), 9842–9857 (2023).
27. Barbosa, L. C., Machado, G. C., Heringer, M. & Ferrer, V. P. Identification of established and novel extracellular matrix components in glioblastoma as targets for angiogenesis and prognosis. *Neurogenetics* **25**(3), 249–262 (2024).
28. Gu, S., Wang, Y., Lei, D. & Zhao, H. Analysis and construction of ceRNA networks reveal 4 mRNAs as potential biomarkers of temozolomide-resistant glioblastomas. *Research Square (preprint)* (2021).
29. Li, Y. et al. Distinct genomic aberrations between low-grade and high-grade gliomas of Chinese patients. *PLoS One* **8**(2), e57168 (2013).
30. Kanehisa, M. & Goto, S. KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res.* **28**(1), 27–30 (2000).
31. Xu, H. et al. Comprehensive molecular characterization of long-term glioblastoma survivors. *Cancer Lett.* **593**, 216938 (2024).
32. Hauffe, L. et al. Eukaryotic translation initiation factor 4E binding protein 1 (EIF4EBP1) expression in glioblastoma is driven by ETS1- and MYBL2-dependent transcriptional activation. *Cell Death Discov.* **8**(1), 91 (2022).
33. Yi, G. Z. et al. Identification of key candidate proteins and pathways associated with temozolomide resistance in glioblastoma based on subcellular proteomics and bioinformatical analysiss. *BioMed Res. Int.* **2018**, 5238760 (2018).
34. Qi, T. et al. Glioma-associated oncogene homolog 1 in breast invasive carcinoma: a comprehensive bioinformatic analysis and experimental validation. *Front. Cell Dev. Biol.* **12**, 1478478 (2024).
35. Sun, Z., Qi, X. & Zhang, Y. Bioinformatics Analysis of the Expression of ATP Binding Cassette Subfamily C Member 3 (ABCC3) in Human Glioma. *Open Med. (Warsaw)* **15**, 107–113 (2020).
36. Hermawan, A. & Putri, H. Systematic analysis of potential targets of the curcumin analog pentagamavunon-1 (PGV-1) in overcoming resistance of glioblastoma cells to bevacizumab. *Saudi Pharm. J.* **29**(11), 1289–1302 (2021).
37. Swiatek-Machado, K. & Kaminska, B. STAT Signaling in Glioma Cells. In *Glioma Signaling* (ed. Barańska, J.) 203–222 (Springer International Publishing, Cham, 2020).

Author contributions

B.B.: Methodology, Software, Validation, Investigation, Writing - Original Draft, Writing - Review & Editing, Visualization; J.J: Methodology, Investigation; M.K.: Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Project administration, Funding acquisition; Y.K.: Conceptualization, Methodology, Validation, Resources, Writing - Original Draft, Writing - Review & Editing, Supervision, Funding acquisition

Funding

This work was supported by the National Science Foundation Major Research Instrumentation (NSF MRI) (Grant#:2117941), the National Research Foundation of Korea (NRF) (NRF-2021R1I1A3048029), and the MSIT (Ministry of Science and ICT) under the ICAN (ICT Challenge and Advanced Network of HRD) support program (IITP-2024-RS-2022-00156409) supervised by the IITP (Institute for Information & Communications Technology Planning & Evaluation) in South Korea.

Declarations

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-026-35273-3.

Correspondence and requests for materials should be addressed to M.K. or Y.K.