BioData Mining

METHODOLOGY Open Access

Sparse generalized linear model with L_0 approximation for feature selection and prediction with big omics data

Zhenqiu Liu[1*], Fengzhu Sun[2] and Dermot P. McGovern[3]

*Correspondence: liuzx@cshs.org
[1] Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles 90048, CA, USA
Full list of author information is available at the end of the article

Abstract

Background: Feature selection and prediction are the most important tasks for big data mining. The common strategies for feature selection in big data mining are L_1, SCAD and MC+. However, none of the existing algorithms optimizes L_0, which penalizes the number of nonzero features directly.

Results: In this paper, we develop a novel sparse generalized linear model (GLM) with L_0 approximation for feature selection and prediction with big omics data. The proposed approach approximate the L_0 optimization directly. Even though the original L_0 problem is non-convex, the problem is approximated by sequential convex optimizations with the proposed algorithm. The proposed method is easy to implement with only several lines of code. Novel adaptive ridge algorithms (L_0ADRIDGE) for L_0 penalized GLM with ultra high dimensional big data are developed. The proposed approach outperforms the other cutting edge regularization methods including SCAD and MC+ in simulations. When it is applied to integrated analysis of mRNA, microRNA, and methylation data from TCGA ovarian cancer, multilevel gene signatures associated with suboptimal debulking are identified simultaneously. The biological significance and potential clinical importance of those genes are further explored.

Conclusions: The developed Software L_0ADRIDGE in MATLAB is available at https://github.com/liuzqx/L0adridge.

Keywords: Sparse modeling, L_0 penalty, Big data mining, Multi-omics data, GLM, Classification, Suboptimal debulking

Background

Integrating multilevel molecular and clinical data to design preventive, diagnostic, and therapeutic solutions that are individually tailored to each patient's requirements is the ultimate goal of precision medicine. However, the huge number of features makes it neither practical nor feasible to predict clinical outcomes with all omics features directly. Thus, selecting a small subset of informative features (biomarkers) to conduct association studies and clinical predictions has become an important step toward effective big data mining. Statistical tests or univariate correlation analysis for feature selection ignore the interacting relationship among genes. To evaluate the predictive power of the features, one appealing approach for feature selection is L_0 regularized sparse modeling, which penalizes the number of nonzero features directly. L_0 is known as the most essential sparsity measure and has nice theoretical properties. However, it is computational impossible to perform an exhaustive search when analyzing omics data sets with millions of features. L_0 penalized optimization is known to be NP-hard in general (Lin et al. 2010).

One common strategy for feature selection is to replace the non-convex L_0 with the L_1 norm. L_1 is a convex relaxation and loose approximation of L_0. Although L_1 penalized sparse models [1] can be solved efficiently, the estimators with L_1 are penalized too much and asymptotically biased. In addition, L_1 inclines to select more spurious features than necessary, and may not always choose the true model consistently [2]. Theoretically, L_1 never outperforms L_0 by a constant [3]. Depending on the location of true optimum, L_1 may perform much worse than L_0 [4, 5]. As a result, the convex relaxation techniques have been shown to be suboptimal in many cases [6]. More recent approaches aimed to reduce bias and overcome discontinuity include the non-convex SCAD [7] and MC+ [8]. However, none of the existing algorithms directly approximate the L_0 optimization problem. Either SCAD or MC+ has been rarely used for feature selection in big data analytics because of their computational intensity with multiple tuning parameters. On the other hand, recent research works including ours show that sparse regression models with L_0 penalty (local solution) outperforms L_1 (global solution) by a substantial margin [5, 9–11].

Debulking cytoreductive surgery is a standard treatment for ovarian cancer. The goal of debulking is to remove as much visible cancer as possible. However, if tumor nodules have invaded vital organs, surgeons may not be able to remove them without compromising the patient's life. Leaving tumor nodules larger than 1 cm is defined as suboptimal debulking (cytoreduction). It has been shown that suboptimal debulking is associated with reduced chemosensitivity and poor survival in ovarian cancer. Biomarkers derived from multi-omics data may help physicians decide which patients should undergo surgery and which should be treated with chemotherapy first [12–14]. Identifying biomarkers from multi-omics data has been an exciting but challenging task. Sparse modeling is one of the important approaches for simultaneous phenotype prediction and biomarker identification. In this paper, we propose a L_0 penalized generalized linear regression (GLM) for feature selection and prediction. Adaptive ridge algorithm (L_0ADRIDGE) is developed to approximate L_0 penalized GLM with sequential convex optimization and is efficient in handling ultra high-dimensional omics data. The proposed method outperforms other cutting-edge convex and non-convex penalties including L_1, SCAD and MC+ with simulations. When applied to the important suboptimal debulking prediction problem in ovarian cancer, the proposed approach identifies multilevel molecular signatures through mining methylation, microRNA and mRNA expression data jointly from TCGA. The identified molecular signatures are further evaluated using public databases.

Materials and methods

Given an input $X_{N\times P}$, where $N \ll P$, and output Y, we have a generalized linear model with canonical link in the following form:

$$E(Y|X) = \mu = G(\theta), \quad \text{and} \quad \theta = X\beta,$$

where G is a canonical link function. Different link functions lead to different models. For instance, a logit link function leads to logistic regression, while an exponential link function leads to Poisson regression.

L_0 penalized GLM

The distribution of Y in GLM is assumed to be from the exponential families with the following probability (density) function:

$$f(Y,\theta,\phi) = \exp\left\{\frac{Y\theta - B(\theta)}{A(\phi)} + C(Y,\phi)\right\},$$

where ϕ is a dispersion parameter, and different functions $A(*)$, $B(*)$ and $C(*)$ are for different distributions Y [15]. The corresponding mean and variance are:

$$E(Y) = \mu = B'(\theta), \text{and } \ Var(Y) = V(\mu)A(\phi) = B''(\theta)A(\phi),$$

where $V(\mu) = B''(\theta)$. Let $Y = [Y_1, \ldots, Y_N]^t$, $X = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]^t$, and $\mu = [\mu_1, \ldots, \mu_N]^t$, so $\mu_i = G(\theta_i) = G\left(\mathbf{x}_i^t\beta\right)$ and $\theta_i = \mathbf{x}_i^t\beta$. The log-likelihood of Y is

$$\begin{aligned} L(Y,\mu,\phi) &= \sum_{i=1}^{N} \log f(Y_i,\theta_i,\phi) \\ &= \sum_{i=1}^{N}\left\{\frac{Y_i\theta_i - B(\theta_i)}{A(\phi)} - C(Y_i,\phi)\right\}. \end{aligned}$$

Dropping the constants $A(\phi)$, and $C(Y_i,\phi)$, we have the simplified log likelihood as follows:

$$L(Y,\mu) = \sum_{i=1}^{N}\{Y_i\theta_i - B(\theta_i)\}.$$

Hence, L_0 penalized error function to minimize is

$$\begin{aligned} \underset{\beta}{\text{argmin}} E &= \underset{\beta}{\text{argmin}}\left\{-L(Y,\mu) + \frac{\lambda}{2}|\beta|_0\right\} \\ &= \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}[B(\theta_i) - Y_i\theta_i] + \frac{\lambda}{2}|\beta|_0\right\}, \end{aligned} \tag{1}$$

where $|\beta|_0 = \sum_{j=1}^{P} I(\beta_j \neq 0)$ is the number of nonzero elements in β, $\mu_i = G(\theta_i)$ and $\theta_i = \mathbf{x}_i^t\beta$. If we define $\frac{0}{0} = 0$, then $|\beta|_0 = \sum_j I(\beta_j \neq 0) = \sum_j \frac{\beta_j^2}{\beta_j^2}$. Equation (2) is equivalent to

$$\begin{aligned} \underset{\beta}{\text{argmin}} E &= \underset{\beta}{\text{argmin}}\{-L(Y,\mu) + \lambda|\beta|_0\} \\ &= \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}[B(\theta_i) - Y_i\theta_i] + \frac{\lambda}{2}\sum_{j=1}^{P}\frac{\beta_j^2}{\beta_j^2}\right\}, \end{aligned} \tag{2}$$

which is equivalent to the following system:

$$\begin{aligned} \underset{\beta}{\text{argmin}} E &= \underset{\beta}{\text{argmin}}\{-L(Y,\mu) + \lambda|\beta|_0\} \\ &= \underset{\beta}{\text{argmin}}\left\{\sum_{i=1}^{N}[B(\theta_i) - Y_i\theta_i] + \frac{\lambda}{2}\sum_{j=1}^{P}\frac{\beta_j^2}{\eta_j^2}\right\}, \\ \eta &= \beta. \end{aligned} \tag{3}$$

Given η and $\theta_i = \mathbf{x}_i^t \beta$, the derivative of E w.r.t. β is

$$\begin{aligned}\nabla E &= \sum_{i=1}^{N} \left[B'(\theta_i) - Y_i\right] \frac{\partial \theta_i}{\partial \beta} + \lambda\beta \oslash \eta^2 \\ &= \sum_{i=1}^{N} \left[B'(\theta_i) - Y_i\right] \mathbf{x}_i + \lambda\beta \oslash \eta^2,\end{aligned}$$

where $\oslash$ indicates element-wise division. The Hessian matrix is

$$H(\beta) = \sum_{i=1}^{N} B''(\theta_i)\mathbf{x}^t\mathbf{x} + \lambda \oslash \eta^2.$$

Let

$$D = \begin{bmatrix} \eta_1^2 & 0 & \dots & 0 \\ 0 & \eta_2^2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \eta_P^2 \end{bmatrix}, \text{ and } V = \begin{bmatrix} V_1 & 0 & \dots & 0 \\ 0 & V_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & V_P \end{bmatrix},$$

where $V_i = V(\mu_i) = B''(\theta_i) = G'(\theta_i)$, $i = 1, \ldots, N$, and let $\tilde{Y}$ $=[Y_1 - B'(\theta_1), \ldots, Y_N - B'(\theta_N)]^t =[Y_1 - \mu_1, \ldots, Y_N - \mu_N]^t$, we have

$$\begin{aligned}\nabla E &= -D^{-1}(DX^t\tilde{Y} - \lambda\beta), \\ H(\beta) &= D^{-1}(DX^tVX + \lambda I).\end{aligned} \tag{4}$$

The Newton-Raphson iteration for β is

$$\begin{aligned}\beta^{new} &= \beta^{old} - \left\{H\left(\beta^{old}\right)\right\}^{-1} \nabla E \\ &= \beta^{old} + \left(DX^tVX + \lambda I\right)^{-1} \left(DX^t\tilde{Y} - \lambda\beta^{old}\right) \\ &= \left(DX^tVX + \lambda I\right)^{-1} \left[DX^tVX\beta^{old} + DX^t\tilde{Y}\right] \\ &= \left(DX^tVX + \lambda I\right)^{-1} DX^t \left[VX\beta^{old} + \tilde{Y}\right].\end{aligned}$$

Let $Z = VX\beta^{old} + \tilde{Y}$, we have

$$\begin{aligned}\beta^{new} &= \left(DX^tVX + \lambda I\right)^{-1} DX^tZ, \\ \eta &= \beta^{old} = \beta^{new}.\end{aligned} \tag{5}$$

Different link functions will lead to different regression models as shown as in Table 1.

Other GLMs such as negative binomial, gamma, and inverse Gaussian can be implemented accordingly with a different $V(\mu)$. When dealing with big data problems with $N \ll P$,where N is the number of samples and P is the number of parameters, the inverse of a $P \times P$ matrix is time-consuming and computational challenging. We proposed an

Table 1 Link functions for linear, logistic and Poisson regression models in GLM, where different models have different $A(*)$, $B(*)$, and $C(*)$

GLM models	$B(\theta)$	$\mu(\theta) = B'(\theta)$	Link $\theta(\mu)$	$V(\mu) = B''(\theta)$
Linear regression	$\theta^2/2$	θ	Identity	1
Logistic regression	$\log(1 + e^{\theta})$	$\frac{1}{1+e^{-\theta}}$	logit	$\mu(1 - \mu)$
Poisson regression	$\exp(\theta)$	$\exp(\theta)$	log	μ

efficient algorithm to calculate the inverse of a much smaller $N \times N$ matrix as follows (Liu et al. 2015):

$$\left(DX^t VX + \lambda I_{P\times P}\right)^{-1} DX^t = DX^t \left(VXDX^t + \lambda I_{N\times N}\right)^{-1}.$$

So that when $N \ll P$, we have a much efficient estimation:

$$\begin{aligned} &\beta^{new} = DX^t \left(VXDX^t + \lambda I\right)^{-1} Z, \\ &\quad \eta = \beta^{old} = \beta^{new}. \end{aligned} \tag{6}$$

The adaptive ridge algorithm (L_0ADRIDGEA) is implemented in MATLAB are as follows:

The L_0ADRIDGE Algorithm:

```
Given a $\lambda > 0$, $\varepsilon = 1e - 6$,
and training data $\{X, \mathbf{y}\}$,
Initializing $\beta^{new} = rand(P, 1)/100$,
While 1,
    $\eta = \beta^{old} = \beta^{new}$, and $D = diag\left(\eta_1^2, \ldots, \eta_P^2\right)$.
    $\theta = X\beta^{old}$, $\mu(\theta) = G(\theta) = B'(\theta)$, and $\tilde{Y} = Y - \mu(\theta)$.
    $V = diag[\,(B''(\theta_1), \ldots, B''(\theta_N)]$, and $Z = VX\beta^{old} + \tilde{Y}$
    If $N \geq P$, $\beta^{new} = \left(DX^t VX + \lambda I\right)^{-1} DX^t Z$,
    Else, $\beta^{new} = DX^t \left(VXDX^t + \lambda I\right)^{-1} Z$.
    if $||\beta^{new} - \eta|| < \varepsilon$, Break; End
End
```

The algorithm is easy to implement and very efficient for either small sample size and large dimension or large sample size and small dimension big data problem. The regularized parameter λ can be determined either by cross-validation or by AIC and BIC with $\lambda = 2$ and $\lambda = \log(N)$, respectively. We further discuss that the proposed method is a L_0 approximation and converges to L_0 when the number of iterations $m \to \infty$.

Algorithm justification: Given a high-dimensional big feature matrix $X_{N\times P}$ ($N \ll P$) and a threshold γ for the coefficient estimates, L_0 rejects all the coefficient estimates below γ to 0 and keeps the large coefficients unchanged. This is the same as defining a binary vector $s = [\ldots, 1, 0, \ldots, 1]^t$, with the value of 0 or 1 for each feature, where $s_j = 1$ if the coefficient estimate for that feature is above the threshold γ, and 0 otherwise. Let $S = \text{diag}(s)$ be a matrix with s on its diagonal, we have the selected feature matrix $X_S = XS$. We can build the standard models with the matrix X_S, if we know s in advance. For instance, we can estimate the coefficients of a GLM with L_2 regulation given X_S and Y with

$$\beta^{new} = (X_S^t VX_S + \lambda I)^{-1} X_S^t Z = (X_S^t VX + \lambda I)^{-1} X_S^t Z = (SX^t VX + \lambda I)^{-1} SX^t Z, \tag{7}$$

where $Z = VX\beta^{old} + \tilde{Y}$, $\tilde{Y} = [\,Y_1 - \mu_1, \ldots, Y_N - \mu_N]^t$, and $X_S^t VX_S = SX^t VXS = SX^t VX$ because of the special structure of matrix S. It is guaranteed that the estimate is 0 for feature j with $s_j = 0$. However, in reality we do not know s. Estimating both s and θ is an NP-hard problem, since we need to solve a mixed-integer optimization problem. Comparing Eq. (7) with Eq. (5), $\beta^{new} = (DX^t VX + \lambda I)^{-1} DX^t Z$, it is clear that S is replaced by D and a binary s_j is approximated by a continuous η_j^2 in proposed algorithm. Therefore, the proposed method is a L_0 approximation.

Recall the iterative system in Eq. (3), note that each feature is penalized by a different penalty, which is inversely proportional to the squared magnitude of that parameter estimator η_j. i.e.,

$$\lambda_j = \frac{\lambda}{2\eta_j^2}, \quad \text{and} \quad \eta_j = \beta_j.$$

Smaller β_j will lead to larger λ_j. A tiny β_j, will become smaller and λ_j will be getting larger in each iteration of L_0ADRIDGE algorithm. $\beta_j \rightarrow 0$, and $\lambda_j \rightarrow \infty$. On the other hand, a larger β_j will lead a finite λ_j, and nonzero β_j, when the number of iteration goes to ∞. The solution of L_0ADRIDGE will converge to that of Eq. (7), because the effect of nonzero η_j will be canceled out in Eq. (5). Note that our proposed methods will find a sparse solution with a large number of iterations and small ε, even though the solution of L_2 regularized modeling is not sparse. Small parameters (β_js) become smaller at each iteration and will eventually go to zero (below the machine ϵ). We can also set a parameter to 0 if it is below predefined $\varepsilon = 1e - 6$ to speed up the convergence of the algorithm.

Results

Simulations

Poisson Regression: Our first simulation was used to evaluate the performance of our method for high dimensional Poisson regression. The data was generated from Poisson distribution with different sample sizes (N) and dimensions (P). However, only features 1, 5, 10 and the constant term are used to generate the Poisson counts with $[\beta_0, \beta_1, \beta_5, \beta_{10}]$ =$[1, 0.5, 0.5, 0.4]$. The count Y is generated with $Y = Poisson(\mu)$, where mean $\mu = \exp(\beta X)$. The proposed method is compared with the glmnet ([16] and SparseReg package [17, 18]. glmnet and SparseReg implemented the elastic net, SCAD, and MC+ penalties with an efficient path algorithm. We compare the performance of our approach with L_1 (glmnet), SCAD and MC+ using the popular BIC ($\lambda = \log(N)$) criteria. Our L_0ADRIDGE is compared to the glmnet for L_1 and SparseReg for both SCAD and MC+. The results of different methods are presented in Table 2.

Table 2 shows that our L_0ADRIDGE consistently achieved the best performance with BIC and different sample sizes and dimensions. With BIC, although MC+ has the lowest square root of mean squared error (rMSE), and fits the data better, L_0ADRIDGE achieves the least absolute bias $|\hat{\beta} - \beta|$, highest percentage of identified true model (PTM), and lowest false discovery rate (FDR) under different simulation settings. The average number of selected features (ANSF) with L_0ADRIDGE is also closest to the true number 4. Particularly, L_0ADRIDGE found 100% true model with the lowest average absolute bias (0.086) under the dimension of $P = 10,000$ and sample size of $N = 500$, indicating that the proposed approach is efficient under extra-high dimensional setting. Another interesting finding is that the square root of mean squared errors and absolute biases with L_0ADRIDGE did not vary much across different simulation setting, indicating the robustness of the proposed approach. Moreover, L_0ADRIDGE with BIC is slightly faster than different routines implemented in glmnet and SparseReg in computational time. Finally, BIC apparently is not a good model selection criteria for L_1, SCAD and MC+. More features are selected than necessary. A larger λ is needed for selecting the correct model. We reported the results with a larger λ on Additional file 1: Table S1, and demonstrated

Table 2 Performance of different GLM methods for Poisson regression over 100 simulations, where values in the parenthesis are the standard deviations, and ANSF: Average number of selected features; rMSE: Average square root of mean squared error; $|\hat{\beta} - \beta| = \sum_i |\hat{\beta}_i - \beta_i|$: average absolute bias when comparing true and estimated parameters

	PMS	glmnet	SparseReg		L_0ADRIDGE
		L_1	SCAD	MC+	
	rMSE	1.10(±.091)	1.090(±.092)	**1.087**(±**.091**)	1.937(±.222)
N =100	$\lvert\hat{\beta} - \beta\rvert$	1.755(±.274)	1.754(±.275)	1.737 ± .273)	**0.222**(±**.116**)
P =100	ANSF	43.03(±3.52)	43.07(±3.57)	42.06(±3.51)	**3.99**(±**.100**)
	PTM	0%	0%	0%	**99**%
	FDR	90.6%	90.6%	90.6%	**0**%
	rMSE	0.503(±.017)	0.502(±.017)	**0.501**(±**.018**)	2.108(±.359)
N =100	$\lvert\hat{\beta} - \beta\rvert$	2.671(±.421)	2.673(±.425)	2.821 ± 2.012)	**0.424**(±**.350**)
$P = 10^3$	ANSF	75.47(±5.61)	75.82(±5.71)	75.14(±8.69)	**3.610**(±**.601**)
	PTM	0%	0%	0%	**64**%
	FDR	94.7%	94.7%	94.6%	**2.4**%
	rMSE	**0.271**(±**.004**)	0.272(±.012)	0.275(±.025)	1.916(±.081)
N =500	$\lvert\hat{\beta} - \beta\rvert$	5.845(±.280)	6.185(±2.359)	5.807 ± .273)	**0.086**(±**.033**)
$P = 10^4$	ANSF	465.6(±14.1)	475.1(±15.5)	463.6(±13.9)	**4.000**(±**.000**)
	PTM	0%	0%	0%	**100**%
	FDR	99.1%	99.2%	99.1%	**0**%

PMS: Performance Measures. PTM: Percentage of true models. FDR: False discovery rate. The values in boldface indicate the best performance

that both SCAD and MC+ can achieve a much smaller FDR, but a larger absolute bias and rMSE.

Logistic regression: The logistic regression data was generated with the coefficients of $[\beta_1, \beta_5, \beta_{10}] = [0.5, 0.5, -0.4]$, respectively, and the remaining coefficients were set to zero. The score $z = X\beta + \varepsilon$, where ε is the random noise with the signal to noise ratio of 4. Then, the probability y is generated from the logistic function $y = 1/(1 + e^{-z})$. Note that y is the true probability instead of binary (1/0) in this simulation. Unlike the previous example, the optimal values of λ in this simulation were selected with the standard 5-fold cross-validation. We divided the λ from $\lambda_{\min} = 1e - 4$, to $\lambda_{\max}$ into 100 equal intervals in log-scale, then chose the optimal λ with the smallest test error. The simulation was also repeated 100 times. The computational results were reported in Table 3. The values in the parenthesis are the positive/negative standard deviation.

Table 3 shows that L_0ADRIDGE outperforms L_1, SCAD and MC+ with a substantial margin under the 5-fold cross-validation. Cross-validation is a standard tool for parameter selection in machine learning. L_0ADRIDGE achieved the smallest test square root of mean squared error, least absolute biases, the lowest FDR, and highest percentages of identified true models The average number of selected features are 3.33 and 3.41 for the dimensions of 100 and 1000, respectively, which are the closest to the true number of features 3. In contrary, L_1, SCAD and MC+ selected unnecessary features. L_1 on average identified 17.1 and 50.92 features, and SCAD selected 18.35 and 73.03 features on average for the dimensions of 100 and 1000, respectively, while MC+ performed slightly better, choosing 10.41 and 24.8 features for the dimensions of 100 and 1000, respectively. More impressively, out of 100 simulations, L_0ADRIDGE identified the true model 81 and 80 times with different dimensions, while L_1 and SCAD could not find the true model

Table 3 Performance of different GLM methods for logistic regression over 100 simulations, where ANSF: Average number of selected features; trMSE: Test Average square root of mean squared error; $|\hat{\beta} - \beta| = \sum_i |\hat{\beta}_i - \beta_i|$: average absolute bias when comparing true and estimated parameters

	PMS	L_1	SparseReg		L_0ADRIDGE
			SCAD	MC+	
	trMSE	0.0474(±.0035)	0.0469(±.0039)	0.0456(±.0042)	**0.0434**(±.**0028**)
N =100	$\lvert\hat{\beta} - \beta\rvert$	0.2984(±.1262)	0.3129(±.1249)	0.1625(±.0752)	**0.0682**(±.**0416**)
P =100	ANSF	17.10(±9.32)	18.35(±10.185)	10.410(±6.174)	**3.330**(±.**779**)
	PTM	0%	0%	2%	**81**%
	FDR	77.7%	78.4%	62.4%	**6.6**%
	trMSE	0.0517(±.0045)	0.0496(±.0046)	0.0468(±.0045)	**0.0434**(±.**0030**)
N =100	$\lvert\hat{\beta} - \beta\rvert$	0.5968(±.2599)	0.6465(±.2205)	0.2818(±.1030)	**0.0754** ± .**0600**)
$P = 1000$	ANSF	50.92(±39.974)	73.030(±40.792)	24.80(±13.314)	**3.41**(±**1.065**)
	PTM	0%	0%	0%	**80**%
	FDR	90.5%	93%	83.9%	**7.3**%

PMS: Performance Measures. PTM: Percentage of true models. FDR: False discovery rate. The values in boldface indicate the best performance

once, and MC+ only identified the true model 2 times for the dimension of 100, indicating the super performance of L_0ADRIDGE under cross-validation. Finally, L_0ADRIDGE is robust. The test square root of mean squared error and other performance measures did not vary much when the dimension increased from 100 to 1000. It is worth noting that our proposed method performs well with the popular statistical model selection criteria such as BIC and cross-validation. Other popular methods such as L_1, SCAD, and MC+ select more features than necessary with such criteria. Therefore, many popular packages including the commercial MATLAB usually choose a larger λ one standard deviation above the minimum test error with cross-validation, which is arbitrary and leads to larger bias. To overcome such bias in parameter estimation, some packages re-estimate the parameters with the selected features and standard GLM model. Unlike these methods, our proposed method performed much better without any postprocessing. Finally, the algorithm is very robust with different initialization. With $N = 100$, $P = 1000$ and 100 times of different randomized initialization, we achieved the trMSE of 0.437(±.003), average absolute bias of 0.0763(±.07), ANSF of 3.39(±1.154), PTM of 85% and FDR of 6.9%, which is quite similar to the results with a fixed initialization.

TCGA ovarian cancer data

The Cancer Genomic Atlas (TCCA) has generated a large amount of next generation sequencing and other omics data for ovarian adenocarcinoma (OC). In this study, we conducted integrated analysis of RNA-seq, miRNA expression, promoter methylation, and debulking status data from 367 OC patients. There are 342 microRNAs, 13,911 mRNA expression (in FPKM), and 21,985 promoter methylation values available. We first normalized different omics data and screened the debulking associated microRNA, mRNAs, and methylation promoters with the P-values of less than 0.01 with the training data only. Based on the central dogma of biology, suboptimal debulking is associated with microRNA expression, gene expression, and DNA methylation; gene expression is a function of microRNA expression and DNA methylation; and microRNA expression is regulated by DNA methylation. L_0 Logistic regression was used for suboptimal debulking prediction, while L_0 penalized Poisson regression was used for gene expression and microRNA expression prediction with FPKM. FPKM, representing fragments per kilobase of exon per million fragments mapped, measures the normalized read counts for RNA-seq. Three-fold cross validation was used for gene selection and validation. We reported the gene signatures with the best predicted area under the ROC curves (AUCs). Molecular signatures that are directly or indirectly associated with suboptimal debulking are shown in Fig. 1.

Figure 1 indicates that there are 16 gene signatures including 7 mRNAs and 9 epigenetic markers directly associated with debulking status. Even though there is no microRNA directly associated with debulking, eight microRNA signatures are indirectly associated with debulking through their association with mRNA signatures. Moreover, there are additional 18 epigenetic markers indirectly associated with debulking. The 7 mRNAs directly associated with debulking are EIF3D, PPP1R7, ADA, HSD17B1, SRBD1, ZNF621, and BARX1, where EIF3D, PPP1R7, BARX1 and ZNF621 have positive correlations and the other 3 genes have negative correlations with suboptimal debulking. Among the 7 mRNAs, ADA (Adenosine Deaminase) is a well-studied gene in ovarian neoplasms. ADA levels were found to be significantly higher in patients with ovarian cancers as compared with benign ovarian tumors [19]. ADA has been regarded as a potential biomarker for diagnosis and an agent for the treatment of ovarian cancer [20]. Other mRNAs such as BARX1, EIF3D, PPP1R7, and HSD17B1 are also known to be associated with different cancers or other diseases. At the microRNA level, there are 8 microRNAs indirectly associated with debulking including mir-183, let-7b, mir-9-1, mir-377, mir-202, mir-758,

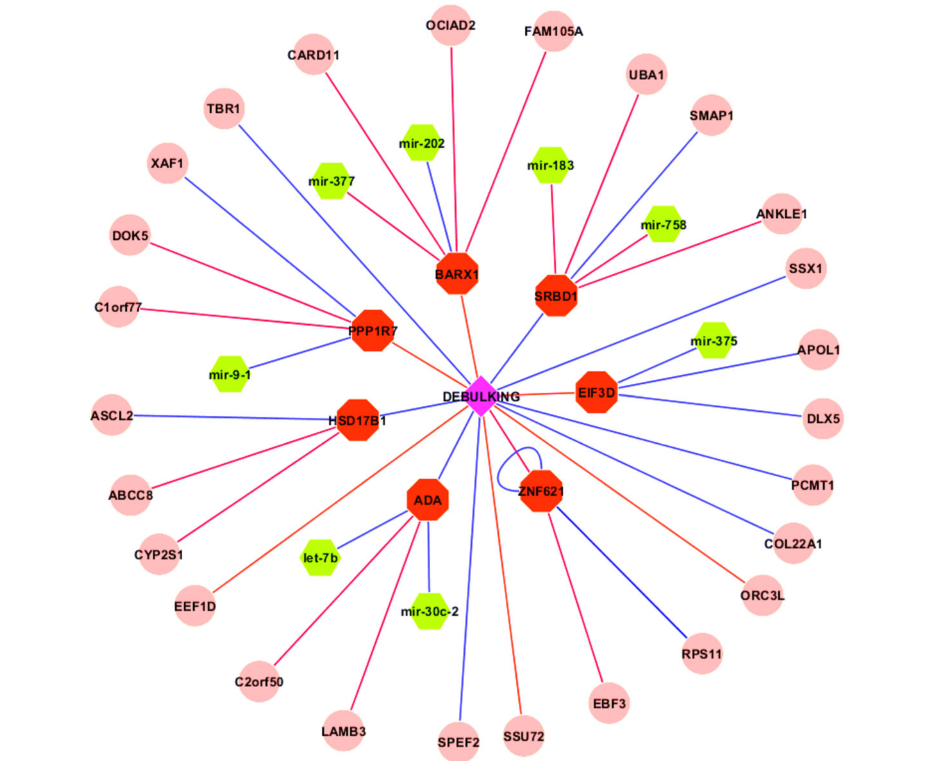

Fig. 1 Gene signatures associated with suboptimal debulking, where nodes in red: mRNA signatures; nodes in green: microRNA signatures; nodes in pink: methylation signatures, and edges in red: positive partial correlation; edges in blue: negative partial correlation

mir-375, and mir-30c-2. While let-7b, mir-30c-2, and mir-377 are positively correlated with suboptimal debulking through mRNAs ADA and BARX1 indirectly, the other 5 microRNAs have indirectly negative correlations with suboptimal debulking. Seven of eight microRNAs except for mir-758 are known to be associated with ovarian cancer. Particularly, let-7b is known to be an unfavorable prognostic biomarker and predict of molecular and clinical subclasses in high-grade serous ovarian carcinoma, and it may also be useful for discriminating between controls and patients with serous ovarian cancer [21, 22]. Mir-183 is known to be associated with multiple cancers. It regulates target oncogene (Tiam1), and reduce the migration, invasion and viability of ovarian cancer cells [23]. Finally, at the DNA level, nine epigenetically modified genes directly associated with debulking are SSX1, TBR1, ZNF621, ORC3L, COL22A1, SPEF2, SSU72, EEF1D, and ZNF621, where EEF1D, SSU72, and ORC3L are positively associated with suboptimal debulking, while 6 other epigenetic genes are negatively correlated with suboptimal debulking. In addition, 18 other epigenetic genes indirectly associated with debulking may also have biological implications. Finally, integration of multi-omic data increases the prediction power substantially. Besides analyzing three types of omics data together, we performed the same three-fold cross validation for gene expression, methylation, and microRNA expression separately. The AUC curves are in Fig. 2.

Figure 2 shows that the best predicted AUC over 100 simulations for integrated data is 0.88, while the best predictive AUCs for gene expression, methylation, and microRNA over 100 simulations are 0.81, 0.84, and 0.76, respectively. The AUC with integrated data achieved the highest AUC, indicating the importance of multi-omics data mining. Genes selected with mRNA, microRNA, and methylations separately are reported in the supplementary document. In addition, we also compare the selected features and the same number of top genes identified with statistical test. The results are reported on Additional file 1: Table S2, and demonstrate that although individual genes are more statistically significant, combination of a panel of genes with standard logistic regression has less predictive power and test AUC (0.79).

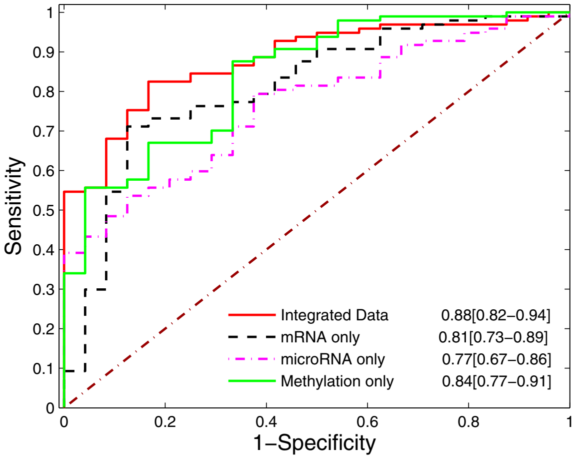

Fig. 2 Predictive AUCs for integrated data, mRNA expression only, microRNA expression only, and methylation only

Conclusions

Biomarkers from multi-omics data may predict disease status and help physicians to make clinical decisions. L_0 based GLM, which directly penalizes the number of nonzero parameters, has nice theoretical properties and leads to essential sparsity for biomarker discovery. Optimizing the L_0 regularization is a crucial, but difficult problem. We have developed an adaptive ridge algorithm (L_0ADRIDGE) for approximating L_0 penalized GLM. The algorithm is easy to implement and efficient for problems with either an ultra-high dimension and small sample size, or a low-dimension and large sample size. It outperforms the other cutting edge regularization methods including L_1, SCAD and MC+ through simulations. When applied to the integration of multilevel omics data from TCGA and the prediction of suboptimal debulking from ovarian cancer, it can identify a panel of gene signatures achieving the best prediction power. We also demonstrate that prediction power of a model with multi-omics data increases substantially, when comparing with a model with one omics data, indicating the importance of big data mining.

Additional file

Additional file 1: Table S1. Performance of different GLM methods for Poisson regression over 100 simulations, where values in the parenthesis are the standard deviations, and ANSF: Average number of selected features; rMSE: Average square root of mean squared error; $|\hat{\beta} - \beta| = \sum_i |\hat{\beta} - \beta_i|$: average absolute bias when comparing true and estimated parameters. PMS: Performance Measures. PTM: Percentage of true models. FDR: False discovery rates. The L0ADRIDGE is compared to the best performance chosen from $\lambda = 0.9\lambda_{\mathrm{max}}$ and $\lambda = 0.5\lambda_{max}$ with both for SCAD and MC+. Table S2. The comparison of performance of the our sparse modeling approach and the top genes selected with Student's t-test. The results demonstrate that although each gene is more statistically significant with statistical test, the combination of the panel of genes has less predictive power and test AUC with standard logistic regression and three-fold cross valida- tion, indicating the collinearity among theses genes. (PDF 86 kb)

Abbreviations

AUC: The area under an ROC curve; GLM: Generalized linear model; L_0ADRIDGE: L_0 adaptive ridge algorithms; MC+: Minimax concave penalty. SCAD: Smoothly clipped absolute deviation; TCGA: The cancer genome atlas

Acknowledgements

This work was partially supported by the DMS-1343506 grant from the National Science Foundation (ZL), and P01 CA098912-10 and UL1 TR0001881-01 grants from NIH. The funders had no role in the preparation of the article.

Funding

DMS-1343506 of NSF (ZL), P01 CA098912-10 and UL1 TR0001881-01.

Availability of data and materials

L_0 ADRIDGE in MATLAB is available at https://github.com/liuzqx/L0adridge.

Authors' contributions

ZL conceptualized and designed method, developed the software, and wrote the manuscript. SF, MDP helped in method design and manuscript writing and revised the manuscript critically. All authors read and approved the final manuscript.

Ethics approval and consent to participate

Not Applicable.

Consent for publication

Not Applicable.

Competing interests

The authors declare that they have no competing interests.

Publisher's Note

Author details

[1]Samuel Oschin Comprehensive Cancer Institute, Cedars-Sinai Medical Center, Los Angeles 90048, CA, USA. [2]Molecular and Computational Biology Program, Department of Biological Sciences, University of Southern California, Los Angeles 90089, CA, USA. [3]Foundation Inflammatory Bowel & Immunobiology Research Institute, Cedars-Sinai Medical Center, Los Angeles, 90048, CA, USA.

Received: 13 April 2017 Accepted: 4 December 2017
Published online: 19 December 2017

References

1. Tibshirani R. Regression shrinkage and selection via the lasso. J Roy Stat Soc B. 1996;58:267–88.
2. Zou H. The adaptive lasso and its oracle properties. J Am Stat Assoc. 2006;101:1418–29.
3. Lin D, Foster D, Ungar L. A risk ratio comparison of l0 and l1 penalized regressions. Tech. rep., University of Pennsylvania; 2010.
4. Kakade S, Shamir O, Sridharan K, Tewari A. Learning exponential families in high dimensions: strong convexity and sparsity. JMLR. 2013;9:381–8.
5. Bahmani S, Raj B, Boufounos P. Greedy sparsity-constrained optimization. J Mach Learn Res. 2013;14(3):807–41.
6. Zhang T. Multi-stage convex relaxation for feature selection. Bernoulli. 2012;19:2153–779.
7. Fan J, Li R. Variable selection via nonconcave penalized likelihood and its oracle properties. J Am Stat Assoc. 2001;96:1348–61.
8. Zhang C. Nearly unbiased variable selection under minimax concave penalty. Ann Stat. 2010;38:894–942.
9. Liu Z, Lin S, Deng N, McGovern D, Piantadosi S. Sparse inverse covariance estimation with L0 Penalty for Network Construction with Omics Data. J Comput Biol. 2016;23(3):192–202.
10. Liu Z, Li G. Efficient Regularized Regression with L_0 Penalty for Variable Selection and Network Construction. Comput Math Methods Med. 2016;2016:3456153.
11. Bahmani S, Boufounos P, Raj B. Learning Model-Based Sparsity via Projected Gradient Descent. IEEE Trans Info Theory. 2016;62(4):2092–9.
12. Riester M, Wei W, Waldron L, Culhane A, Trippa L, Oliva E, Kim S, Michor F, Huttenhower C, Parmigiani G, Birrer M. Risk prediction for late-stage ovarian cancer by meta-analysis of 1525 patient samples. J Natl Cancer Inst. 2014;106(5): Apr 3.
13. Tucker S, Gharpure K, Herbrich S, Unruh A, Nick A, Crane E, Coleman R, Guenthoer J, Dalton H, Wu S, Rupaimoole R, Lopez-Berestein G, Ozpolat B, Ivan C, Hu W, Baggerly K, Sood A. Molecular biomarkers of residual disease after surgical debulking of high-grade serous ovarian cancer. Clin Cancer Res. 2014;20(12):3280–8.
14. Liu Z, Beach J, Agadjanian H, Jia D, Aspuria P, Karlan B, Orsulic S. Suboptimal cytoreduction in ovarian carcinoma is associated with molecular pathways characteristic of increased stromal activation. Gynecol Oncol. 2015;139(3): 394–400.
15. Wood S. Generalized Additive Models: An Introduction with R. New York: Chapman & Hall/CRC; 2006.
16. Friedman J, Hastie T, Tibshirani R. Regularization Paths for Generalized Linear Models via Coordinate Descent. J Stat Softw. 2011;33(1):1–22.
17. Zhou H, Lange K. A path algorithm for constrained estimation. J Comput Graph Stat. 2013;22(2):261–83.
18. Zhou H, Wu Y. A generic path algorithm for regularized statistical estimation. J Am Stat Assoc. 2014;109(506):686–99.
19. Urunsak I, Gulec U, Paydas S, Seydaoglu G, Guzel A, Vardar M. Adenosine deaminase activity in patients with ovarian neoplasms. Arch Gynecol Obstet. 2012;286(1):155–9.
20. Shirali S, Aghaei M, Shabani M, Fathi M, Sohrabi M, Moeinifard M. Adenosine induces cell cycle arrest and apoptosis via cyclinD1/Cdk4 and Bcl-2/Bax pathways in human ovarian cancer cell line OVCAR-3. Tumour Biol. 2013;34(2):1085–95.
21. Tang Z, Ow G, Thiery J, Ivshina A, Kuznetsov V. Meta-analysis of transcriptome reveals let-7b as an unfavorable prognostic biomarker and predicts molecular and clinical subclasses in high-grade serous ovarian carcinoma. Int J Cancer. 2014;134(2):306–18.
22. Chung Y, Bae H, Song J, Lee J, Lee N, Kim T, Lee K. Detection of microRNA as novel biomarkers of epithelial ovarian cancer from the serum of ovarian cancer patients. Int J Gynecol Cancer. 2013;23(4):673–9.
23. Li J, Liang S, Jin H, Xu C, Ma D, Lu X. Tiam1, negatively regulated by miR-22, miR-183 and miR-31, is involved in migration, invasion and viability of ovarian cancer cells. Oncol Rep. 2012;27(6):1835–42.