Bioinformatics, 37(4), 2021, 514–521
doi: 10.1093/bioinformatics/btaa776
Advance Access Publication Date: 11 September 2020
Original Paper

Gene expression

Incorporating prior knowledge into regularized regression

Chubing Zeng , Duncan Campbell Thomas and Juan Pablo Lewinger*

Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, CA 90033, USA

*To whom correspondence should be addressed.

Associate Editor: Jonathan Wren

Received on March 12, 2020; revised on August 13, 2020; editorial decision on August 24, 2020; accepted on September 1, 2020

Abstract

Motivation: Associated with genomic features like gene expression, methylation and genotypes, used in statistical modeling of health outcomes, there is a rich set of meta-features like functional annotations, pathway information and knowledge from previous studies, that can be used *post hoc* to facilitate the interpretation of a model. However, using this meta-feature information *a priori* rather than *post hoc* can yield improved prediction performance as well as enhanced model interpretation.

Results: We propose a new penalized regression approach that allows *a priori* integration of external meta-features. The method extends LASSO regression by incorporating individualized penalty parameters for each regression coefficient. The penalty parameters are, in turn, modeled as a log-linear function of the meta-features and are estimated from the data using an approximate empirical Bayes approach. Optimization of the marginal likelihood on which the empirical Bayes estimation is performed using a fast and stable majorization–minimization procedure. Through simulations, we show that the proposed regression with individualized penalties can outperform the standard LASSO in terms of both parameters estimation and prediction performance when the external data is informative. We further demonstrate our approach with applications to gene expression studies of bone density and breast cancer.

Availability and implementation: The methods have been implemented in the R package *xtune* freely available for download from https://cran.r-project.org/web/packages/xtune/index.html.

Contact: lewinger@usc.edu

1 Introduction

Predicting outcomes based on genomic biomarkers, such as gene expression, methylation and genotypes, are becoming increasingly important for individualized risk assessment and treatment (Kamel and Al-Amodi, 2017). As an example, consider predicting mortality from breast cancer after surgical treatment based on gene expression profiles (Nuyten *et al.*, 2006). Since genomic studies typically have more available features than subjects, a common approach to develop prediction models based on genomic features is to use regularized regression methods, which can handle high-dimensional data. In addition to regularization, sparsity inducing regression approaches, such as the LASSO can also perform feature selection. In the context of genomic studies, feature selection is critical for yielding interpretable models that provide insight into potential biological mechanisms and which can, in turn, facilitate adoption by practitioners. To enhance model interpretability, it is common to examine features selected in a model in relation to available information about gene function and previous studies. For example, analyses can be conducted to formally assess whether the selected features are enriched in particular metabolic pathways or gene ontology annotations. This kind of *post hoc* analysis relating genomic features to existing knowledge about them, hereafter referred to as genomic meta-features, can provide valuable biological insight and validation for a prediction model. In this article, we propose a new approach that exploits genomic meta-features a priori rather than *post hoc*, to improve prediction performance and feature selection, and to enhance the interpretability of models developed using penalized regression.

To incorporate meta-features into the model building process, our main idea is to use the penalty parameters in penalized regression as the instrument. Commonly, most regularized regression methods apply a single global penalty parameter to all regression coefficients, effectively treating all features or predictors equally in the model building process. This may result in over-shrinking of important coefficients and under-shrinking of unimportant ones, with a corresponding loss in prediction ability. We extend the standard LASSO regression to allow for penalty terms that depend on external meta-features. Specifically, rather than using a single penalty parameter to control the amount of shrinkage for all regression coefficients, our model allows each coefficient to have its own individual penalty parameter, which is, in turn, modeled as a log linear function of the meta-features. Some examples of prior knowledge include (i) gene function annotation from databases like the Gene Ontology Project (Ashburner *et al.*, 2000); (ii) gene–disease co-occurrence scores from text-mining biomedical abstracts (Pletscher-Frankild *et al.*, 2014; Rouillard *et al.*, 2016); (iii) deleterious somatic mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al.*, 2011). We focus on the LASSO penalty because of its widespread use but address potential extensions to other penalties in the discussion.

Previously, other variants of LASSO regression have been introduced to allow either coefficient-specific penalties or multiple tuning parameters. Yuan and Lin (2006) proposes group LASSO that extends LASSO to grouped predictors. The adaptive LASSO proposed by Zou (2006) also adjusts the penalty individually for each coefficient by using a vector of adaptive weights. The group LASSO applies only to grouping variables, and the adaptive LASSO uses pre-specified weights obtained from the initial estimate of the coefficients using the same data as the data used for regression. Neither of these approaches incorporates a general set of meta-features. Boulesteix *et al.* (2017) proposed the integrative LASSO with penalty factors method that assigns different penalty factors to different data modalities such as gene expression, methylation and copy number. They use cross-validation to choose the penalty parameters based on prediction performance. In practice, the number of different modalities they can allow is up to four, due to computational bottle-neck.

In addition to those above, several other methods have been proposed previously to make use of prior knowledge of the features. Tharmaratnam *et al.* (2016) suggests using biologic knowledge to derive a set of features that could replace the gene set by chosen by LASSO with minimal loss in predictive power. However, 'experts' are required to assess the importance of each gene. Tai and Pan (2007) partitions the features into groups and shrink the features of different groups by different magnitudes. Shrinkage for groups is considered fixed and arbitrary but is not data-dependent, and the use of an external dataset only provides information on the grouping of predictors. van de Wiel *et al.* (2016) proposes an adaptive group-regularized ridge regression that accounts for group structure as the group LASSO and allows group-specific penalties. However, the method requires partitioning of the features into different groups, and the external information is only used for guiding the partition. The approach proposed by Bergersen *et al.* (2011) is also in the form of the adaptive LASSO. Instead of constructing weights from the same data X and Y, they use the Spearman correlation coefficients or the ridge regression coefficients between the features X and external information as the penalty weights.

Our approach is distinguished from these methods in that (i) we adopt an empirical Bayes approach to estimate the hyperparameters instead of cross-validation, which allows us to estimate feature-specific tuning parameters; (ii) the magnitude of the penalty terms are modeled as a log-linear function of the external information and are estimated from a 'second-level' regression; (iii) our approach is not restricted to meta-features that define feature groupings but can handle meta-features of any type including quantitative ones.

2 Materials and methods

2.1 Model specification

We start by considering a standard linear regression model $Y = X\boldsymbol{\beta}+\boldsymbol{\epsilon}$, where Y is the vector of observed responses for n subjects, X is an $n \times p$ matrix of genomic features, and $\boldsymbol{\epsilon}$ represents independent errors with zero expectation and a common variance σ^2. The LASSO regression shrinks the regression coefficients by imposing a L_1 penalty on the sum of absolute value of regression coefficients (Tibshirani, 1996). The objective function of LASSO is:

$$\min_{\beta\in\mathbb{R}^p}\{||Y - X\beta||_2^2 + \lambda\sum_{j=1}^{p}|\beta_j|\}. \tag{1}$$

The tuning parameter λ controls the strength of regularization: with a larger value of λ, more regression coefficients β are shrunk toward zero. The choice of λ determines feature selection and is key to achieve good model performance.

Our method extends the objective function of LASSO to allow each coefficient β_j to have its own individual penalty parameter λ_j, which is in turn modeled as a log-linear function of the meta-features. The objective function of external information tuned (xtune) LASSO is:

$$\begin{gathered}\min_{\beta\in\mathbb{R}^p}\{||Y - X\beta||_2^2 + \sum_{j=1}^{p}\lambda_j|\beta_j|\}\\ \boldsymbol{\lambda} = \mathrm{e}^{Z\boldsymbol{\alpha}},\end{gathered} \tag{2}$$

where $\boldsymbol{\lambda} = (\lambda_1, \dots, \lambda_\mathrm{p})$ is a vector of tuning parameters, Z is the meta-feature data of dimension $p \times q$, $\boldsymbol{\alpha}$ is the second level coefficients of dimension $q \times 1$ that links external information Z to individual penalties $\boldsymbol{\lambda}$. Note that the standard LASSO and the adaptive LASSO are two special cases of our model, corresponding to the case where Z is a single column of 1 s and Z the identity matrix of dimension $p \times p$, respectively.

Fitting of xtune LASSO consists of two steps: (i) choose the penalty parameters vector $\boldsymbol{\lambda}$ and (ii) estimate the regression coefficients $\boldsymbol{\beta}$ given $\boldsymbol{\lambda}$. The second step is easy to implement using standard softwares, such as the *glmnet* package in the R language (Friedman *et al.*, 2010). For the first step, cross-validation is commonly used to select the single penalty parameter for standard lasso. However, with potentially a large number of penalty parameters to tune, cross-validation is not feasible for xtune LASSO. On the other hand, most penalized regression approaches have a Bayesian interpretation which provides a natural way to allow multiple tuning parameters and incorporate external information.

LASSO regression can be equivalently formulated as a Bayesian or random effects model where the coefficients are modeled with a double exponential (a.k.a. Laplace) prior distribution condition on σ^2 (Tibshirani, 1996). Assuming that the distribution of the response variable Y conditional on the regression coefficients $\boldsymbol{\beta}$ and σ^2 follows a normal distribution, the LASSO coefficient estimates that solves Equation (1) can be equivalently characterized as the posterior mode, or maximum a posteriori (MAP) in a Bayesian model with the a double exponential prior distribution for $\boldsymbol{\beta}$ conditional on σ^2. The regression coefficients $\boldsymbol{\beta}$ that minimize the xtune LASSO objective function in Equation (2) is equivalent to the MAP estimator $\boldsymbol{\beta}_{MAP}$ under the Bayesian formulation shown in Equation (3), which is conditional on σ^2. Therefore, we select the penalty parameter vector $\boldsymbol{\lambda}$ for xtune LASSO in Equation (2) by estimating $\boldsymbol{\lambda}$ as hyperparameters under the Bayesian formulation of xtune LASSO in Equation (3).

$$\begin{gathered}Y|X,\boldsymbol{\beta}\sim\mathcal{N}(\mathrm{X}\boldsymbol{\beta},\sigma^2\mathrm{I_n})\beta_\mathrm{j}|\sigma^2\sim \text{ Double Exponential}(0,\frac{\lambda_\mathrm{j}}{2\sigma^2}),\\ \forall j = 1, 2, .., p\\ \boldsymbol{\lambda} = \mathrm{e}^{Z\boldsymbol{\alpha}}.\end{gathered} \tag{3}$$

The variance of the double exponential prior in the Bayesian formation of xtune LASSO in Equation (3) is $2/\left(\frac{\lambda_j}{2\sigma^2}\right)^2$. Thus, λ_j controls how far away from zero β_j can vary, i.e. the amount of shrinkage for coefficient β_j. A large λ_j implies a small β_j in magnitude. The hyperparameter α_k models in turn how the degree of shrinkage λ_j varies with the value of meta-feature k for feature k, z_{kj}.

Specifically, because $\lambda_j = e^{\alpha_0}e^{\sum_{k=1}^{q}\alpha_k z_{jk}}$, $\lambda_0 = e^{\alpha_0}$ can be interpreted as the overall level of shrinkage across all regression coefficients. The hyperparameters $\alpha_{k,k>0}$ control the individual degree of shrinkage, above or below the overall level given by λ_0, that is applied to coefficient β_j. A positive α_k indicates that features with positive (negative) values of z_k should be shrunk more (less) relative to the overall level of shrinkage λ_0. Similarly, a negative α_k indicates that features with positive (negative) values of z_k should be shrunk less (more) relative to the overall level of shrinkage λ_0. A zero α_k indicates that meta-feature k does not affect the shrinkage of the regression coefficients. Thus, the hyperparameters α_k determine the importance of meta-feature k in determining the magnitude of the regression coefficients. The population variance parameter σ^2 can be estimated from the data or given a point-mass prior (Li and Lin, 2010). In this article, we estimate σ^2 using the method proposed by Reid *et al.* (2016) and assume it is 'set in advance'. Park and Casella (2008) and Li and Lin (2010) used a specification that assigned a non-informative prior for σ^2 and used a Gibbs sampler based on full conditional distributions to obtain regression coefficients estimates. Our method is different in that, rather than extending the model to include Bayesian inference over the hyperparameters, we use an empirical Bayes approach that maximizes the log-likelihood of the hyperparameters. That is, the hyperparameters $\boldsymbol{\alpha}$ are estimated from the data by first marginalizing over the coefficients β and then performing what is commonly referred to as empirical Bayes, evidence maximization or type-II maximum likelihood (Tipping, 2001).

2.2 Empirical Bayes parameter tuning

The empirical Bayes of hyperparameters $\boldsymbol{\alpha}$ (hence $\boldsymbol{\lambda}$) is obtained by maximizing the marginal likelihood calculated by integrating out the random effects $\boldsymbol{\beta}$ from the joint distribution of Y and $\boldsymbol{\beta}$. The marginal likelihood of $\boldsymbol{\alpha}$ is given by:

$$L(\boldsymbol{\alpha}) = \int_{\mathbb{R}^p} \mathrm{L}(\mathrm{Y}|\mathrm{X}, \boldsymbol{\beta}) \mathrm{f}(\boldsymbol{\beta}|\boldsymbol{\lambda}(\boldsymbol{\alpha})) \mathrm{d}\boldsymbol{\beta}$$

$$= \int_{\mathbb{R}^p} \mathcal{N}(Y|X\boldsymbol{\beta}, \sigma^2 \mathrm{I_n}) \text{Double Exponential}(\boldsymbol{\beta}|0, \boldsymbol{\alpha}) \mathrm{d}\boldsymbol{\beta}$$

$$= \int_{\mathbb{R}^p} \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - X_i\boldsymbol{\beta})^2}{2\sigma^2}} \prod_{j=1}^{p} \frac{\exp(Z_j\alpha)}{4\sigma^2} e^{-\frac{\exp(Z_j\alpha)}{2\sigma^2}|\beta_j|}, \tag{4}$$

When X is not orthogonal, the marginal likelihood resulting from the p-dimensional integral in Equation (4) does not have a usable closed-form solution. Foster *et al.* (2008) proposes using a Laplace approximation to the integral which has simple close-form solution. Motivated by their approach, we propose a simpler new approximation which uses a normal distribution with the same prior variance to approximate the double exponential distribution. That is, we use a normal distribution $\beta_j \sim \mathcal{N}(0, \frac{2}{\tau_j^2})$ to approximate the double exponential distribution $\beta_j \sim$ Double Exponential(τ_i), yields closed-form solution for the approximated marginal likelihood:

$$\mathrm{Y}|\mathrm{X}, \boldsymbol{\beta} \sim \mathcal{N}(\mathrm{X}\boldsymbol{\beta}, \mathrm{I_n}\sigma^2)$$

$$\boldsymbol{\beta} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{V^{-1}}), \tag{5}$$

where $V = \text{diag}(\eta_1, \ldots \eta_p) = \text{diag}(\left(\frac{\boldsymbol{\lambda}}{2\sigma^2}\right)^2/2)$.

Therefore, the approximate log marginal negative likelihood of $\boldsymbol{\alpha}$ integrating out $\boldsymbol{\beta}$ is:

$$-\ell(\boldsymbol{\alpha}) = \log|\mathrm{C}_\alpha| + \mathrm{Y}^\mathrm{T}\mathrm{C}_\alpha^{-1}\mathrm{Y} \tag{6}$$

where $C_\alpha = \sigma^2 I + XV^{-1}X^T$. The approximated log likelihood (6) is then maximized to obtain $\boldsymbol{\alpha}$ estimates. Once $\boldsymbol{\alpha}$ known, hence the penalty parameters vector $\boldsymbol{\lambda}$ known, the *glmnet* package in the R language is used to implement the LASSO with given penalty vector.

2.3 Marginal likelihood maximization

The objective function given by the negative marginal log-likelihood (Equation 6) is non-convex, making it intrinsically a challenging problem. We note that the approximated model (Equation 5) is closely related to the model specification of the Automatic Relevance Determination (ARD) (MacKay, 1992; Neal, 1995; Tipping, 2001) method widely used in the field of signal processing. Wipf and Nagarajan (2008, 2010) described a Majorization Minimization (MM) procedure that uses a reformulation of ARD to optimize the non-convex optimization function by solving a series of easier re-weighted L_1 problem. Motivated by their idea, we propose an iterative re-weighted L_2 optimization algorithm described in detail below. Note that this non-convex optimization problem is a special case of the difference of convex functions (DC) problem (Le Thi and Pham Dinh, 2018).

The log-determinate term $\log|C_\alpha|$ is a concave function in $\boldsymbol{\alpha}$ (Boyd and Vandenberghe, 2004). A majorization function of it is its slope at the current value α of $\log|C_\alpha|$:

$$\boldsymbol{\theta} = \nabla_{\boldsymbol{\alpha}} \log|\mathrm{C}_\alpha| = \text{diag}[\mathrm{X}^\mathrm{T}\mathrm{C}_\alpha^{-1}\mathrm{X}]. \tag{7}$$

The $Y^T C_\alpha^{-1} Y$ term in Equation (6) is convex. Therefore, at current value of $\boldsymbol{\alpha}$, the majorization function for $-\ell(\boldsymbol{\alpha})$ is $\ell_{\boldsymbol{\theta}}(\boldsymbol{\alpha}) \triangleq \boldsymbol{\theta}^\mathrm{T}\frac{1}{\eta} + \mathrm{Y}^\mathrm{T}\mathrm{C}_\alpha^{-1}\mathrm{Y}$. Given $\boldsymbol{\theta}$, $\boldsymbol{\alpha}$ is updated by:

$$\boldsymbol{\alpha} \leftarrow \underset{\alpha}{\text{argmin}}\, \ell_\theta(\alpha) \triangleq \theta^\mathrm{T}\frac{1}{\eta} + \mathrm{Y}^\mathrm{T}\mathrm{C}_\alpha^{-1}\mathrm{Y}. \tag{8}$$

Although the objective function (Equation 8) is convex, it is slow to optimize in practice. We use one more MM procedure for optimizing (Equation 8). The data dependent term $Y^T C_\alpha^{-1} Y$ can be re-expressed as:

$$\mathrm{Y}^T\mathrm{C}_\alpha^{-1}\mathrm{Y} = \min_{\delta} \frac{1}{\sigma^2}||\mathrm{Y} - X\delta||^2 + \sum_j \delta_j^2 \eta_j. \tag{9}$$

We therefore introduce another auxiliary term δ, the upper-bounding auxiliary function for $\ell_{\boldsymbol{\theta}}(\boldsymbol{\alpha})$ is:

$$\ell_\theta(\alpha, \delta) \triangleq \theta^T \frac{1}{\eta} + \sum_j \eta_j \delta_j^2 + \frac{1}{\sigma^2}||Y - X\delta||^2 \geq \ell_\theta(\alpha). \tag{10}$$

The $\boldsymbol{\alpha}$ value that minimizes Equation (8) can be estimated by iteratively updating $\boldsymbol{\delta}$ and $\boldsymbol{\alpha}$ in Equation (10). For any $\boldsymbol{\delta}$, $\boldsymbol{\alpha}$ is estimated by minimizing

$$\theta^T \frac{1}{\eta} + \sum_j \eta_j \delta_j^2. \tag{11}$$

Given $\boldsymbol{\alpha}$, $\boldsymbol{\delta}$ is updated by:

$$\delta \leftarrow \underset{\delta}{\text{argmin}}\, ||Y - X\delta||^2 + \sum_j \eta_j \delta_j^2. \tag{12}$$

Equation (12) has a weighted convex L_2 regularized cost function, and it can be optimized efficiently using *glmnet*. The iterative reweighted L_2 algorithm has the schema summarized in Algorithm 1. Simulations we conducted with continuous instead of binary meta-features yielded very similar results (data not shown).

Algorithm 1: Optimization algorithm

```
Step 1: Initialize $\boldsymbol{\alpha}$ with $\boldsymbol{\alpha_0}$ (e.g, $\alpha_i = 0, \forall i = 1, 2, ..., q$)
while not converge do
    Step 2: Given $\boldsymbol{\alpha}$, update $\theta$ by:
        $\theta^{new} \leftarrow \text{diag}[X^T C_{\boldsymbol{\alpha}}^{-1} X]$
    ;
    Step 3: Given $\theta$, update $\boldsymbol{\alpha}$ by the following inner loop:
        • Step 3.1: initialize $\boldsymbol{\alpha}$ with estimates from last iteration
        • Step 3.2: given $\boldsymbol{\alpha}$, update $\boldsymbol{\delta}$ by:
            $\boldsymbol{\delta} \leftarrow \underset{\boldsymbol{\delta}}{\text{argmin}}\, ||Y - X\boldsymbol{\delta}||^2 + \sum_j \eta_j \delta_j^2$
        • Step 3.3: given $\boldsymbol{\delta}$, update $\boldsymbol{\alpha}$ by:
            $\boldsymbol{\alpha} \leftarrow \underset{\boldsymbol{\alpha}}{\text{argmin}}\, \boldsymbol{\theta}^T \frac{1}{\boldsymbol{\eta}} + \sum_j \eta_j \delta_j^2$
        • Step 3.4: iterate Step 3.2 and Step 3.3 until converge
end
```

Once the hyperparameters $\boldsymbol{\alpha}$ are estimated, and therefore the penalties are known, the LASSO regression coefficients can be obtained using standard LASSO software (e.g. glmnet).

2.4 Extension to linear discriminant analysis for classification

So far, we have been focused on LASSO linear model where the response variable is continuous. Here, following the scheme proposed by Mai *et al.* (2012) that builds high dimensional linear discriminate analysis (LDA) upon sparse penalized linear regression, we extend the xtune LASSO model to the framework of LDA with a binary response variable $Y \in \{1, 2\}$.

The LDA model assumes that X is normally distributed within each class, i.e.

$$X|(Y = k) \sim \mathcal{N}(\mu_k, \Sigma)$$
$$\Pr(Y = k) = \pi_k, \ \ k = 1, 2$$

where μ_k is the mean of X within class k and Σ is the common within-class covariance matrix. We adopt the following procedure proposed by Mai *et al.* (2012) to predict the class of Y based on xtune LASSO linear penalized regression:

Step 1. Let $y_i = -\frac{n_1}{n}$ if $Y_i = 1$, and $y_i = \frac{n}{n_2}$ if $Y_2 = 2$.
Step 2. Compute the solution to a penalized least squares problem:

$$(\hat{\beta}^{\text{LDA}}, \hat{\beta_0}^{\text{LDA}}) = \text{argmin}_{\beta,\beta_0}\, n^{-1}\sum_{i=1}^{n}(y_i - \beta_0 - x_i^T\beta)^2 + \sum_{j=1}^{p}P_\lambda(\beta_j) \tag{13}$$

Step 3. Estimate the LDA model on the reduced data $\{Y_i, X_i^T\hat{\beta}^{\text{LDA}}\}_{i=1}^{n}$. Assign observation x to class 2 if

$$x\{-(\hat{\mu}_1 + \hat{\mu}_2)/2\}^T\hat{\beta}^{\text{LDA}} + (\hat{\beta}^{\text{LDA}})^T\hat{\Sigma}\hat{\beta}^{\text{LDA}} \{(\hat{\mu}_1 - \hat{\mu}_2)^T\hat{\beta}^{\text{LDA}}\}^{-1}\log(n_2/n_1) > 0, \tag{14}$$

where $\hat{\mu}_1, n_1, \hat{\mu}_2, n_2$ are the sample mean vector and sample size within class 1 and class 2, $\hat{\Sigma}$ is the sample covariance matrix, $P_\lambda(.)$ is a generic sparsity-inducing regularization term, such as the single-penalty L_1 norm, Elastic-Net or adaptive L_1 norm in our case. We first solve Equation (13) using xtune LASSO, then predict response variable class by Equation (14).

3 Results

3.1 Simulation studies

3.1.1 Simulation setting

We performed a simulation study to evaluate the performance of the xtune LASSO under a range of scenarios obtained by varying the following key simulation parameters:

1. The true ability of the features to predict the outcome as measured by the signal to noise ratio (SNR) defined as $\text{Var}(X\boldsymbol{\beta})/\sigma^2$.
2. The informativeness of the external metadata controlled by the number of non-zero hyperparameters α.
3. The number of predictor features p.
4. The sparsity level captured by the number of non-zero true regression coefficients $\boldsymbol{\beta}$.
5. The overall magnitude of the non-zero entries of the true regression coefficient vector $\boldsymbol{\beta}$ controlled by α_0.
6. The degree of correlation between features ρ.

The simulation data was generated according to the following steps: (i) set the parameter $\boldsymbol{\alpha}$ controlling the informativeness of the external metadata; (ii) generate the external metadata matrix Z; (iii) generate the vector of regression coefficients $\boldsymbol{\beta}$ based on $\boldsymbol{\alpha}$ and Z; (iv) generate the feature data X; (v) generate outcome Y based on β, X and an independent drawn random error $\boldsymbol{\epsilon}$.

The external metadata matrix Z of dimension $p \times q$ was generated to have 0–1 entries indicating a grouping of features. Specifically, entry Z_{jk} indicates whether the jth feature belongs to group k (1) or not (0). The entries were independently generated to be zero or one with probability 0.2 and 0.8, respectively. A consequence of how Z was generated is that features can belong to more than one group, i.e. the groups can overlap. A column of 1 s was appended to Z corresponding to the second level intercept α_0, which controls the overall amount of shrinkage; the higher α_0; the smaller the regression coefficients.

The true regression coefficient β_j were generated by sampling from a double exponential distribution with local parameter $\mu = 0$ and scale parameter $b = \exp(Z_j\boldsymbol{\alpha})$. Parameter δ controls the sparsity of the final $\boldsymbol{\beta}$ vector. The $[p^\delta]$ largest βs in magnitude were retained, while all smaller entries were set to be zero. The feature matrix X was simulated by sampling from a multivariate normal distribution with mean vector zero and covariance matrix $\Sigma_{i,j} = \rho^{|i-j|}$.

Finally, for each simulation replicate, an independent error term ϵ was sampled from a normal distribution with mean 0 and a variance corresponding to the pre-set SNR. The outcome was generated as $Y = X\boldsymbol{\beta} + \epsilon$. Each simulated data is split into a training set ($n = 200$) and a test set ($n = 1000$). The performance of the standard LASSO, adaptive LASSO and xtune LASSO were compared by the prediction R^2 computed on the test using the model fitted in the training set. The large test set guarantees that the generalization/test error was accurately estimated. Penalty parameter tuning for the standard LASSO was performed by 10-fold cross-validation. Implementation of adaptive LASSO utilizes the *adalasso()* function in the *parcor* R package. One hundred replicates were generated for each scenario.

We considered the following six scenarios varying each of the main simulation parameters described above. The simulation results are summarized inFigure 1.

1. $SNR = 1, 2, 3$
2. $p = 500, 1000, 3000$
3. External data informativeness: low, medium and high. We fixed the sub-set of the $\boldsymbol{\alpha} = (\alpha_0, -1, -0.78, -0.56, -0.33, -0.11, 0.11, 0.33, 0.56, 0.78, 1)$.The low external information is simulation from $\alpha_{\text{low}} = (\alpha, \underbrace{0, \ldots, 0}_{60})$. The medium external information is simulated from $\alpha_{\text{medium}} = (\alpha, \underbrace{0, \ldots, 0}_{20})$, and the high external information is simulated from $\boldsymbol{\alpha}_{\text{high}} = \boldsymbol{\alpha}$. The idea is that non-informative external information has many noise variables.
4. $\alpha_0 = 1, 3, 5$
5. $\delta = 0.3, 0.5, 0.7$
6. Different correlation magnitude $\rho = 0.3, 0.6, 0.9$

3.1.2 Simulation results

Figure 1a shows box plots (across simulation replicates) of the prediction R^2 for the standard, adaptive and xtune LASSO varying SNR. As SNR increases, the prediction performance increases for all methods, but the xtune LASSO has a better prediction accuracy across all levels of SNR considered.

When the p/n ratio gets higher, all three methods have decreased performance. However, xtune LASSO has a slower decreasing rate. When the p/n ratio is very high, we see that the performance of standard LASSO decreased dramatically, and xtune LASSO has significantly better prediction performance than standard LASSO (Fig. 1b). The R^2 for standard LASSO remain the same across different level of external data informativeness. The performance of xtune LASSO decreases with external information of lower informativeness (Fig. 1c).

Figure 1d shows the effect of decreased $\boldsymbol{\beta}$ sparsity. The higher the value of δ, the less sparse the model and the more non-zero regression coefficients. The value of xtune LASSO becomes more apparent when the model is less sparse. With $\delta = 0.5$, 31 out of 1000 features are non-zero. All three methods have decreased performance when the model has more non-zero features, but xtune LASSO has a slower decreasing rate. From Figure 1e, the overall amount of shrinkage seems to have little effect on the performance of all three methods. Notice that all three methods have a higher prediction

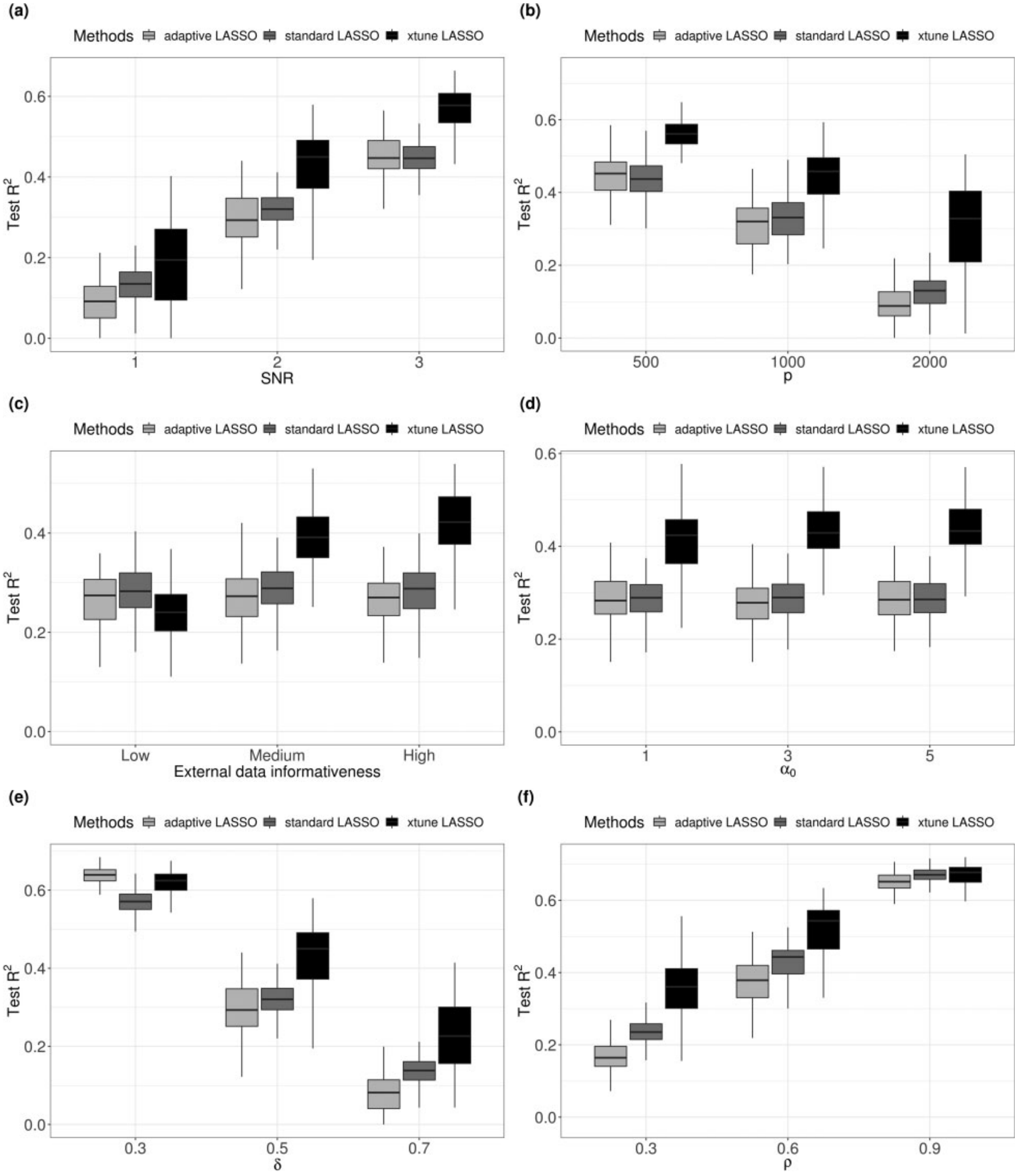

Fig. 1. Simulation results. Subplot (**a**): SNR $= 1, 2, 3$, with $n = 200$, $p = 1000$, $q = 10$, $\alpha_0 = 3$, $\delta = 0.5$, $\rho = 0.2$. Subplot (**b**): $p = 500$, 1000, 3000, with $n = 200$, SNR $= 2$, $q = 10$, $\alpha_0 = 3$, $\delta = 0.5$, $\rho = 0.2$. Subplot (**c**): $q = 10$, 30, 50, with $n = 200$, $p = 1000$, SNR $= 2$, $\alpha_0 = 3$, $\delta = 0.5$, $\rho = 0.2$. Subplot (**d**): regression coefficients sparsity $\delta = 0.3, 0.5, 0.7$, with $n = 200$, $p = 1000$, SNR $= 2$, $q = 10$, $\alpha_0 = 3$, $\rho = 0.2$. Subplot (**e**): overall penalty magnitude $\alpha_0 = 1, 3, 5$, $n = 200$, $p = 1000$, SNR $= 2$, $q = 10$, $\delta = 0.5$, $\rho = 0.2$. Subplot (**f**): $\rho = 0.3, 0.6, 0.9$, with $n = 200$, $p = 1000$, SNR $= 2$, $q = 10$, $\alpha_0 = 3$, $\delta = 0.5$

ability when the signal features (non-zero) features are correlated with each other (Fig. 1f). The improved prediction of xtune LASSO over standard LASSO becomes smaller as the correlation between features increases. LASSO is known to have a decreased ability in variable selection when the features are highly correlated. It tends to select one variable from each highly correlated group and ignoring the remaining ones. Hebiri and Lederer (2013) studied the influence of correlation on LASSO prediction and suggested that correlation in features is not problematic for LASSO prediction; the prediction errors are mostly smaller for the correlated settings in experimental studies.

In summary, our simulation results showed that the LASSO with individual penalties informed by meta-features can outperform the standard LASSO in terms of prediction when (i) the meta-features are informative for the regression effect sizes, (ii) the true model is less sparse and (iii) the SNR is relatively high.

3.2 Applications

We exemplify the method's performance on real data by considering two applications to clinical outcomes prediction using gene expression data.

3.2.1 Bone density data

In the first example, bone biopsies from 84 women were profiled using an expression microarray to study the relationship between bone mass density (BMD) and gene expression. The goal is to predict the total bone density based on the gene expression profiles. Bone density was measured by the hip T-score derived from biopsies, with a higher score indicating higher bone density. The data contains 22 815 gene expression features and were normalized using the RMA method as described in (Tharmaratnam *et al.*, 2016). Gene expression levels were analyzed on a logarithmic-2 scale. The bone density dataset is publicly available from the European Bioinformatics Institute (EMBL-EBI) Array Expression repository ID E-MEXP-1618.

The external information consists of four covariates. The first covariate uses insights from a previous study that identified eight genes highly associated with bone density variation (Reppe *et al.*, 2010). A binary external covariate indicates whether each gene is one of the eight genes identified or not. The second to fourth external covariates are evidence scores related to empirical *P*-values with continuous values that indicates the strength of functional annotations. They are extracted from the dbGAP Gene-Trait Associations dataset, GWASdb SNP-Phenotype Associations dataset and GWAS Catalog SNP-Phenotype Associations dataset (Li *et al.*, 2012; Rouillard *et al.*, 2016; Welter *et al.*, 2013).

To illustrate the advantage of incorporating external information, we compare our proposed method to both standard LASSO and adaptive LASSO. The data were randomly split into a training data consisting of 80% of the observations and a test dataset consisting of 20% of the observations. We fitted the adaptive LASSO, standard LASSO, and our proposed method in the training data and evaluated their prediction performance in the testing data. We repeated 100 random splits of the full data into training and test sets. Figure 2 shows the MSE, R^2 and the number of selected (non-zero) expression features across the 100 splits.

Overall, we see that the externally tuned LASSO has better prediction performance than the standard LASSO while selecting a more parsimonious model. The adaptive LASSO does not perform well in this data example. To gain further insight into the prediction performance results, we examined the penalties applied to the regression coefficients by each of the methods when fitted on the full data. The tuning parameter chosen by standard LASSO using cross-validation is 0.16, while for the xtune LASSO, the estimated tuning parameter is 0.26 for the gene expression features without external information and 0.016 for the expression features with external information, resulting in larger effects estimates for the latter group. The estimated α for the first external covariate is negative, which means the group of eight genes are penalized less than the genes not in this group. Among the identified genes, SOST and DKK1 are involved in the Wnt pathway which is central to bone turnover. ABCA8 and NIPSNAP3B are involved in transporting lipids across membranes and vesicular trafficking (Reppe *et al.*, 2010).

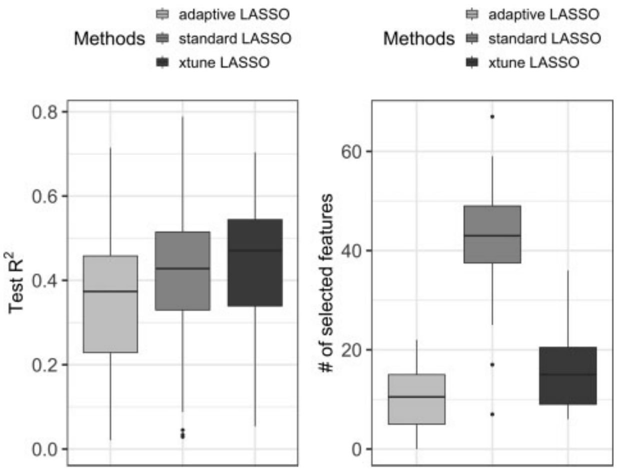

Fig. 2. Compare test R^2 and number of selected covariates of adaptive LASSO, standard LASSO and xtune LASSO using bone density data. The mean test R^2 is 0.27 for adaptive LASSO; 0.38 for the standard LASSO, and 0.43 for xtune LASSO. The mean number of selected covariates is 10 for adaptive LASSO, 43 for standard LASSO and 16 for the xtune LASSO

3.2.2 Breast cancer data

In the second example, we apply the xtune LASSO to a breast cancer dataset. The data is from the Molecular Taxonomy of Breast Cancer International Consortium (METABRIC) cohort (Curtis *et al.*, 2012) (https://ega-archive.org/dacs/EGAC00001000484). 29 476 gene expression profiles and three clinical variables (age at diagnosis, progesterone receptor status and lymph node status) were used for the prediction of five-year survival (survived or died). Patients followed-up for less than five years, with no record of mortality, were excluded from the analysis. We used a subset of the METABRIC data with patients that are Estrogen receptor (ER) positive and human epidermal growth factor receptor 2 (HER2) status negative. The data contains a discovery set of 594 observations and an additional validation set of 564 observations. The models were trained in the discovery set and tested on the validation set.

The external information used for the xtune LASSO model consists of six covariates. The first covariates is a binary variable indicating whether a predictor in *X* is a clinical feature (1) or a gene expression feature (0). The second covariate is a continuous variable with evidence scores for genes co-occurring with the breast cancer in abstracts of biomedical publications from the DISEASES Text-mining Gene–Disease Association Evidence Scores dataset (Pletscher-Frankild *et al.*, 2014). The third covariate has count values for the number of times that each gene is used as one of major prognostic signatures for breast cancer. More specifically, breast cancer multigene prognostic signatures including OncotypeDX (21 genes), MammaPrint (70 genes), PAM50 (50 genes), EndoPredict (12 genes) and Breast Cancer Index (7 genes) have been validated through clinical trials and are recommended for classification, prognosis and prediction in breast cancer. We therefore believe genes that belong to one or more prognostic signatures are potentially more important than other genes. The frequency that each gene belongs to one of breast cancer prognostic signatures is used as the third external covariate. The fourth to sixth external covariates are based on the results of (Cheng *et al.*, 2013), where groups of genes referred to as 'metagenes' that are prognostic in all cancers, including breast cancer were identified. In specific, (Cheng *et al.*, 2013) analyzed six gene expression datasets from different cancer types and present three multi-cancer attractors with strong phenotype associations: a lymphocyte-specific attractor (LYM), a mesenchymal transition attractor strongly associated with tumor stage (MES), and a mitotic chromosomal instability attractor strongly associated with tumor grade (CIN). The LYM, MES and CIN metagenes consist of 169, 134 and 108 genes, respectively. The fourth to sixth external covariates are binary variables to indicate whether or not each gene belong to LYM, MES and CIN respectively.

We compared the xtune LASSO incorporating the meta-features described above, with the standard and the adaptive LASSO. As in the first example, standard LASSO was tuned by repeated 10 fold cross-validation and the adaptive LASSO is implemented using the *adalasso()* function in the *parcor* R package.

Table 1 compares the AUC, the number of selected features, and the computation time for the standard, the adaptive and the xtune LASSO. Figure 3 shows the receiver operating characteristic (ROC) curves for the three methods. The xtune LASSO has the best prediction performance among all three methods. The second level coefficients $\boldsymbol{\alpha}$ estimated by xtune LASSO is $\boldsymbol{\alpha} = (7.23, -4.55, -0.57, -0.13, -1.42, -0.47, -1.38)$. This illustrates how the xtune LASSO can induce differential shrinkage among the features according to their empirical importance. In this

Table 1. Compare AUC, number of selected covariates and computation time (in minutes) for standard LASSO, adaptive LASSO and xtune LASSO

	Standard LASSO	Adaptive LASSO	xtune LASSO
AUC	0.677	0.718	0.767
No. of selected covariates	207	5	13
Computation time	2.26	39.28	11.00

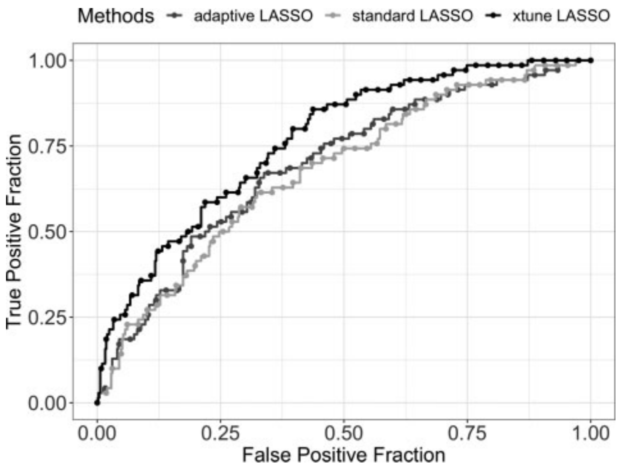

Fig. 3. ROC curves for adaptive LASSO, standard LASSO and xtune LASSO applied to the breast cancer dataset

example, the three clinical variables are given a very small penalty in xtune LASSO and are all estimated to be non-zero. The xtune LASSO shrinks the coefficients corresponding to expression features that do not belong to any metagene (the vast majority) toward zero much more aggressively than the standard LASSO, while it shrinks the clinical variables and the expression features in metagenes LYM and MES much less than the standard LASSO. The standard LASSO shrink the coefficients for the two clinical variables to 0, effectively selecting them out of the model. Adaptive LASSO selected out one of the clinical variables (lymph node status). In agreement with our simulation results, the xtune LASSO also yielded a much more parsimonious model with only 13 selected features, while the standard LASSO selected 207 features. In terms of computation time, both adaptive LASSO (fitted using *adalasso* function) and xtune LASSO take more computation time than the standard LASSO fitted using *glmnet*.

4 Discussion

We introduced xtune LASSO, a new approach implemented in the namesake R package (Zeng and Lewinger, 2019) for integrating external meta-features to inform the tuning of penalty parameters in LASSO regression. We showed through simulation and real data examples that xtune could yield better prediction performance than standard LASSO by incorporating prior knowledge. These findings are consistent with the work of (van de Wiel *et al.*, 2016), which proposed a method for estimating group-specific penalties for ridge regression and showed that the use of prior knowledge could improve prediction performance.

xtune LASSO differs from related methods (Boulesteix *et al.*, 2017; Liu *et al.*, 2018; Pan *et al.*, 2010) in our use of an empirical Bayes approach to estimate the penalty parameters, rather than relying on cross-validation. In the particular case of no external meta-features (i.e. Z is a vector of 1 s), xtune performs empirical Bayes tuning of the single LASSO penalty parameter, providing an alternative to standard tuning by cross-validation. Cross-validation becomes impractical with more than a handful of penalty parameters, while empirical Bayes tuning allows xtune to handle a much larger number of individualized penalties.

The empirical Bayes approach in this article estimates the individual variances of the first-level regression coefficients by maximizing the marginal likelihood. This is also known as Type II maximum likelihood estimation, which has been employed for fitting the relevance vector machine model, a popular technique in the pattern recognition machine learning community. Instead of relying on cross-validation, the 'Bayesian' set-up has distinct advantages in terms of the penalty parameter choice by marginalizing them over the posterior distribution (Bhattacharya *et al.*, 2014). Bayesian penalization methods (Li and Lin, 2010; Park and Casella, 2008) also have employed this connection and use hierarchical models to select the penalty parameter based on sampling strategies. A similar empirical Bayes approach has also been used by the penalized regression methods implemented in the EBglmnet R package (Huang and Liu, 2016). However, these methods do not incorporate information from external covariates, which is the main goal of this article.

An MM algorithm is used to minimize the non-convex objective function (6). While there may be multiple local minima to (6), we did not encounter multi-modality issues in the analyses described in the manuscript. The same estimates are obtained for different initial starting values of $\boldsymbol{\alpha}$. In sensitivity analysis, we also tried directly optimizing the non-convex objective function (6) using standard methods (EM algorithm, gradient descent, L-BFGS algorithm), the same hyperparameter estimates are obtained but the MM algorithm is much faster than the other methods.

In both of our real data application examples, the meta-features are combination of continuous variables and categorical variables that group features into subsets. A categorical meta-feature also arises when the features originate from different data types (e.g. gene expression, methylation, somatic mutations). (Liu *et al.*, 2018) showed that having separate penalty parameters for each data type can yield better prediction performance.

The gain in prediction performance in the proposed model depends to the relevance of the external information that is used to guide the penalization. The external information is relevant or 'informative' if it can help differentiate groups of predictors of different effect size or correlates with the importance of the predictors. Therefore, expert knowledge of the study domain is often crucial. A question that we have not addressed in the article is the how to decide whether the inclusion of an external covariate would be helpful before they are included in the model. In practice, the use of xtune LASSO is more demanding than the use of the standard LASSO as it requires more time and thought on the extraction, processing and interpretation of the external information. On the other hand, however, the process of finding and understanding external information about the predictors can help the investigators better understand the predictors, rather than fitting the model as a 'black box'. We do not seek to claim that our approach always give a better performance as compared to standard LASSO or other related methods. Rather, we provides a way to integrate prior knowledge into the model building process and show that is quite competitive, especially on larger datasets and the meta-features are informative.

Although prediction performance has been our main focus of interest, our results also show that for the range of simulation scenarios we considered and in the two real data applications, xtune tends to yields sparser and, therefore, more interpretable models than standard LASSO regression. However, we did note that when the number of meta-features q relative to the sample size n is large, the $\boldsymbol{\alpha}$ estimates may not be stable. A related limitation is that in its current implementation, xtune does not scale to ultra high dimensional datasets. Typical datasets that xtune LASSO can currently handle have sample size of up to $n \approx 5000$, with $p \approx 50\,000$ features and $q \approx 100$ meta-features. However, we believe that future algorithmic improvements, along with parallel computing, can extend the applicability of xtune to larger datasets and larger numbers of meta-features. To further widen the range of applicability of xtune, we are pursuing extensions to binary (logistic regression) and time to event (Cox regression) outcomes, as well as the incorporation of the Ridge and Elastic-Net penalties in addition to the LASSO.

Funding

This work was supported by the National Cancer Institute at the National Institutes of Health [P01 CA196569, R01 CA140561].

Conflict of Interest: none declared.

References

Ashburner,M. *et al.* (2000) Gene ontology: tool for the unification of biology. The gene ontology consortium. *Nat. Genet.*, **25**, 25–29.

Bergersen,L.C. *et al.* (2011) Weighted lasso with data integration. *Stat. Appl. Genet. Mol. Biol.*, **10.**

Bhattacharya,A. *et al.* (2015) Dirichlet-Laplace priors for optimal shrinkage. *J. Am. Stat. Assoc.*, **110**, 1479-1490.

Boulesteix,A.-L. *et al.* (2017) IPF-LASSO: integrative-penalized regression with penalty factors for prediction based on multi-omics data. *Comput. Math. Methods Med.*, **2017**, 1–14.

Boyd,S. and Vandenberghe,L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.

Cheng,W. *et al.* (2013) Biomolecular events in cancer revealed by attractor metagenes. *PLoS Comput. Biol.*, **9**, e1002920.

Curtis,C. *et al.*; METABRIC Group. (2012) The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**, 346–352.

Forbes,S. *et al.* (2011) Cosmic: mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic Acids Res.*, **39**, D945–50.

Foster,S.D. *et al.* (2008) A random model approach for the LASSO. *Comput. Stat.*, **23**, 217–233.

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1.

Hebiri,M. and Lederer,J. (2013) How correlations influence lasso prediction. *IEEE Trans. Inf. Theory*, **59**, 1846–1854.

Huang,A. and Liu,D. (2016) EBglmnet: a comprehensive r package for sparse generalized linear regression models. *Bioinformatics*, https://academic.oup.com/bioinformatics/advance-article/doi/10.1093/bioinformatics/btw143/2453188.

Kamel,H. and Al-Amodi,H. (2017) Exploitation of gene expression and cancer biomarkers in paving the path to era of personalized medicine. *Genomics Proteomics Bioinf.*, **15, 220-235.**

Le Thi,H.A. and Pham Dinh,T. (2018) DC programming and DCA: thirty years of developments. *Math. Programm.*, **169**, 5–68 .

Li,M. *et al.* (2012) GWASdb: a database for human genetic variants identified by genome-wide association studies. *Nucleic Acids Res.*, **40**, D1047–D1054.

Li,Q. and Lin,N. (2010) The Bayesian elastic net. *Bayesian Anal.*, **5**, 151–170.

Liu,J. *et al.* (2018) Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*, **19**, 369.

MacKay,D.J.C. (1992) Bayesian Interpolation. *Neural Comput.*, **4**, 415–447.

Mai,Q. *et al.* (2012) A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99**, 29–42.

Neal,R.M. (1995) *Bayesian Learning for Neural Networks, Volume 118*. Springer Science and Business Media. https://link.springer.com/bookseries/694.

Nuyten,D. *et al.* (2006) Predicting a local recurrence after breast-conserving therapy by gene expression profiling. *Breast Cancer Res. BCR*, **8**, R62.

Pan,W. *et al.* (2010) Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, **66**, 474–484.

Park,T. and Casella,G. (2008) The Bayesian Lasso. *J. Am. Stat. Assoc.*, **103**, 681–686.

Pletscher-Frankild,S. *et al.* (2014) Diseases: text mining and data integration of disease–gene associations. *Methods (San Diego, Calif.)*, 74, 83-89.

Reid,S. *et al.* (2016) A study of error variance estimation in lasso regression. *Statistica Sinica,* 26, 35–67.

Reppe,S. *et al.* (2010) Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. *Bone*, **46**, 604–612.

Rouillard,A. *et al.* (2016) The harmonizome: a collection of processed datasets gathered to serve and mine knowledge about genes and proteins. *Database*, **2016**, baw100.

Tai,F. and Pan,W. (2007) Incorporating prior knowledge of predictors into penalized classifiers with multiple penalty terms. *Bioinformatics*, **23**, 1775–1782.

Tharmaratnam,K. *et al.* (2016) Tilting the lasso by knowledge-based post-processing. *BMC Bioinformatics*, **17**, 1–9.

Tibshirani,R. (1996) Regression Shrinkage and Selection via the Lasso. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **58**, 267–288.

Tipping,M. (2001) Sparse Bayesian learning and the relevance vector mach. *J. Mach. Learn. Res.*, **1**, 211–244.

van de Wiel,M.A. *et al.* (2016) Better prediction by use of co-data: adaptive group-regularized ridge regression. *Stat. Med.*, **35**, 368–381.

Welter,D. *et al.* (2013) The NHGRI GWAS catalog, a curated resource of SNP-trait associations. *Nucleic Acids Res.*, **42**(Database issue): D1001-D1006.

Wipf,D. and Nagarajan,S. (2008) A new view of automatic relevance determination. *Compute*, **20**, 1625–1632.

Wipf,D. and Nagarajan,S. (2010) Iterative reweighted l1 and l2 methods for finding sparse solutions. *IEEE J. Select. Top. Signal Process.*, **4**, 317–329.

Yuan,M. and Lin,Y. (2006) Model selection and estimation in regression with grouped variables. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, **68**, 49–67.

Zeng,C. and Lewinger,J.P. (2019) *xtune: Regularized Regression with Differential Penalties Integrating External Information.* R package version 1.0.0. https://cran.r-project.org/web/packages/xtune/index.html.

Zou,H. (2006) The adaptive lasso and its oracle properties. *J. Am. Stat. Assoc.*, **101**, 1418–1429.