Incorporating Prior Knowledge into Regularized Regression

Chubing Zeng¹, Duncan Campbell Thomas¹, and Juan Pablo Lewinger^{1,*}

¹Division of Biostatistics, Department of Preventive Medicine, Keck School of Medicine, University of Southern California, Los Angeles, California.

*Corresponding author: lewinger@usc.edu

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 based 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 CRAN.

1 Introduction

Predicting outcomes based on genomic biomarkers such as gene expression, methylation, and genotypes is 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 paper, 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.

Most penalized 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 can result in over-shrinking of important coefficients and under-shrinking of unimportant ones, with a corresponding loss in prediction ability. In genomics and other applications, there is often a great deal of prior knowledge about the features. Examples of prior knowledge include (1) gene function annotation from databases like the Gene Ontology Project (Ashburner *et al.*, 2000); (2) gene-disease co-occurrence scores from text-mining biomedical abstracts (Pletscher-Frankild *et al.*, 2014; Rouillard *et al.*, 2016); (3) deleterious somatic mutations in the Catalogue of Somatic Mutations in Cancer (COSMIC) (Forbes *et al.*, 2010). With the main goal of improving prediction but also to improve feature selection performance, parameter estimation, and model interpretability, we extend the standard LASSO regression to allow for penalty terms that depend on such external meta-features. Specifically, rather than using a single penalty parameter that controls the global amount of shrinkage across all regression coefficients, our model allows each coefficient to have its own individual penalty parameter, which is in turn modeled as a function of the meta-features. 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. Zou (2006) adjusts the penalty differently 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 incorporate 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 per-

formance. In practice, the number of different modalities they can allow is up to 4, 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 genes 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 account for group structure as the group LASSO and allow group-specific penalties. However, the method require 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 hyper-parameters 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 Methods

2.1 Random effect formulation of the LASSO

We start by considering a standard linear regression model $Y = X\beta + \epsilon$, where Y is the vector of observed responses for n subjects, X is an $n \times p$ matrix of genomic features, and ϵ 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:

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

Most penalized regression approaches have a Bayesian interpretation. The 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](#)). In specific, we assume that the distribution of the response variable Y conditional on the regression coefficients β and σ^2 follows a normal distribution. The LASSO coefficient estimates that solves (1) can be equivalently characterized as the posterior mode, or maximum a posteriori (MAP) in a Bayesian model with the following double exponential prior distribution for β conditional on σ^2 :

$$\begin{aligned} Y|X, \boldsymbol{\beta} &\sim N(X\boldsymbol{\beta}, \sigma^2 I_n) \\ f(\boldsymbol{\beta}|\sigma^2) &\sim \text{Double Exponential}(0, \frac{\lambda}{2\sigma^2}) \end{aligned} \quad (2)$$

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

$$\begin{aligned} \min_{\boldsymbol{\beta} \in \mathbb{R}^p} \{ ||Y - X\boldsymbol{\beta}||_2^2 + \lambda_j \sum_{j=1}^p |\beta_j| \} \\ \log \boldsymbol{\lambda} = Z\boldsymbol{\alpha} \end{aligned} \quad (3)$$

where Z is the meta-feature of dimension $p \times q$, the j^{th} column of Z comprise information about feature j . $\boldsymbol{\alpha}$ is the unknown hyperparameter vector that links the individual penalties and the external information.

By using the Bayesian formulation on penalized regression, external covariates can be incorporated via the prior on the model coefficients. Here we allow differential penalization by assigning priors with different scale parameters to regression coefficients. In specific, the double exponential random effect formulation of the standard LASSO in (2) assumes a single common ‘prior’ variance for all β_j ’s. In our extension, we let each β_j have a double exponential random effect distribution but with its own individual hyperparameter λ_j , where the hyperparameter vector $\boldsymbol{\lambda}$ is in turn modeled as a log-linear function of the external information available for feature j , Z .

$$\begin{aligned} Y|X, \boldsymbol{\beta} &\sim \mathcal{N}(X\boldsymbol{\beta}, \sigma^2 I_n) \\ f(\beta_j|\sigma^2) &\sim \text{Double Exponential}(0, \frac{\lambda_j}{2\sigma^2}), \forall j = 1, 2, \dots, p \\ \log \boldsymbol{\lambda} &= Z\boldsymbol{\alpha} \end{aligned} \quad (4)$$

The maximum a posteriori (MAP) estimator under the random effects model in (4), which is conditional on σ^2 , is equivalent to the estimator minimizing the LASSO objective function with external information driven penalization in (3). The population variance parameter σ^2 can be estimated from the data or given a point-mass prior (Li and Lin, 2010). In this paper, we estimate σ^2 using the method proposed by Reid *et al.* (2013) and assume it is ‘set in advance’. Park and Casella (2008) and Li and Lin (2010) used a specification that assigns 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 hyperparameters. That is, the hyperparameters $\boldsymbol{\alpha}$ are estimated from the data by first marginalizing over the coefficients $\boldsymbol{\beta}$ 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

Unlike the standard LASSO, the proposed model has a potentially large number of hyperparameters ($\alpha_{1 \times q}$), so tuning them by cross-validation is not feasible. Instead, we propose an empirical Bayes approach for directly estimating the hyperparameters based on the training data. Specifically, the empirical Bayes of α (hence λ) is obtained by maximizing the marginal likelihood calculated by integrating out the random effects β from the joint distribution of Y and β .

The marginal likelihood of α is given by:

$$\begin{aligned} L(\alpha) &= \int_{\mathbb{R}^p} p(Y|X, \beta) f(\beta|\alpha) d\beta \\ &= \int_{\mathbb{R}^p} \mathcal{N}(Y|X\beta, \sigma^2 I_n) \text{DE}(\beta|0, \alpha) d\beta \\ &= \int_{\mathbb{R}^p} \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(Y_i - X_i\beta)^2}{2\sigma^2}} \prod_{j=1}^p \frac{\exp(Z_j\alpha)}{4\sigma^2} e^{-\frac{\exp(Z_j\alpha)}{2\sigma^2} |\beta_j|} d\beta \end{aligned} \quad (5)$$

When X is not orthogonal, the marginal likelihood resulting from the p -dimensional integral in (5) does not have a usable close-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 $\beta_j \sim \text{Double Exponential}(\tau_j)$, yields closed form solution for the approximated marginal likelihood:

$$\begin{aligned} Y|X, \beta &\sim \mathcal{N}(X\beta, I_n\sigma^2) \\ \beta &\sim \mathcal{N}(0, V^{-1}) \end{aligned} \quad (6)$$

where

$$V = \begin{bmatrix} \eta_1 & & \\ & \ddots & \\ & & \eta_p \end{bmatrix} = \begin{bmatrix} \frac{1}{2\sigma^2} \exp(2Z_1\alpha) & & \\ & \ddots & \\ & & \frac{1}{2\sigma^2} \exp(2Z_p\alpha) \end{bmatrix}$$

Therefore, the approximate log marginal likelihood of α integrating out β is:

$$-\ell(\alpha) = \log |C_\alpha| + Y^T C_\alpha^{-1} Y \quad (7)$$

where $C_\alpha = \sigma^2 I + XV^{-1}X^T$. The approximated log likelihood (7) is then maximized to obtain α estimates. Once α known, hence the penalty parameters vector λ known, the *glmnet* package (Jerome Friedman, Trevor Hastie and Tibshirani, 2010) in the R language is used to implement the LASSO with given penalty vector.

2.3 Optimizing the Objective Function

The objective function (7) is non-convex, making it intrinsically a challenging problem. We note that the approximated model (6) 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, 2014) describe 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).

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

$$\theta = \nabla_\alpha \log |C_\alpha| = \text{diag}[X^T C_\alpha^{-1} X] \quad (8)$$

The $Y^T C_\alpha^{-1} Y$ term in (7) is convex. Therefore, at current value of α , the majorization function for $-\ell(\alpha)$ is $\ell_\theta(\alpha) \triangleq \theta^T \frac{1}{\eta} + Y^T C_\alpha^{-1} Y$. Given θ , α is updated by:

$$\alpha \leftarrow \underset{\alpha}{\text{argmin}} \ell_\theta(\alpha) \triangleq \theta^T \frac{1}{\eta} + Y^T C_\alpha^{-1} Y \quad (9)$$

Although the objective function (9) is convex, it is slow to optimize in practice. We use one more MM procedure for optimizing (9). Note that the data dependent term $Y^T C_\alpha^{-1} Y$ can be re-expressed as:

$$Y^T C_\alpha^{-1} Y = \min_{\delta} \frac{1}{\sigma^2} \|Y - X\delta\|^2 + \sum_j \delta_j^2 \eta_j \quad (10)$$

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

$$\ell_\theta(\alpha, \delta) \triangleq \theta^T \frac{1}{\eta} + \sum_j \eta_j \delta_j^2 + \frac{1}{\sigma^2} \|Y - X\delta\|^2 \geq \ell_\theta(\alpha) \quad (11)$$

The α value that minimizes equation (9) can be estimated by iteratively updating δ and α in equation (11). For any δ , α is estimated by minimizing

$$\theta^T \frac{1}{\eta} + \sum_j \eta_j \delta_j^2 \quad (12)$$

Given α , δ is updated by:

$$\delta \leftarrow \underset{\delta}{\text{argmin}} \|Y - X\delta\|^2 + \sum_j \eta_j \delta_j^2 \quad (13)$$

Equation (13) has a weighted convex L_2 regularized cost function, and it can be optimized efficiently using *glmnet*. In summary, the iterative reweighted L_2 algorithm has the following schema:

Algorithm 1 Optimization algorithm

Step 1: Initialize α with α_0 (e.g, $\alpha_i = 0, \forall i = 1, 2, \dots, q$)

while *not converge* **do**

Step 2: Given α , update θ by:

$$\theta^{new} \leftarrow \text{diag}[X^T C_\alpha^{-1} X]$$

Step 3: Given θ , update α by the following inner loop:

- Step 3.1: initialize α with estimates from last iteration
- Step 3.2: given α , update δ by:

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

- Step 3.3: given δ , update α by:

$$\alpha \leftarrow \underset{\alpha}{\text{argmin}} \theta^T \frac{1}{\eta} + \sum_j \eta_j \delta_j^2$$

- Step 3.4: iterate Step 3.2 and Step 3.3 until converge

end

Once the hyperparameters α are estimated, and therefore the penalties 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.,

$$\begin{aligned} X|Y = k &\sim \mathcal{N}(\mu_k, \Sigma) \\ \Pr(Y = k) &= \pi_k, k = 1, 2 \end{aligned} \tag{14}$$

where μ_k is the mean of X within class k and Σ is the common within class covariance matrix. We adapt 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 $y_i = \frac{n_2}{n}$ if $Y_i = 2$.

Step 2. Compute the solution to a penalized least squares problem:

$$(\hat{\beta}^{LDA}, \hat{\beta}_0^{LDA}) = \underset{\beta, \beta_0}{\operatorname{argmin}} n^{-1} \sum_{i=1}^n (y_i - \beta_0 - x_i^T \beta)^2 + \sum_{j=1}^p P_\lambda(\beta_j) \quad (15)$$

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

$$\{x - (\hat{\mu}_1 + \hat{\mu}_2)/2\}^T \hat{\beta}^{LDA} + (\hat{\beta}^{LDA})^T \hat{\Sigma} \hat{\beta}^{LDA} \{(\hat{\mu}_1 - \hat{\mu}_2)^T \hat{\beta}^{LDA}\}^{-1} \log(n_2/n_1) > 0 \quad (16)$$

where $\hat{\mu}_1, n_1, \hat{\mu}_2, n_2$ are the sample mean vector and sample size within class 1 and class2, respectively. $\hat{\Sigma}$ is the sample covariance matrix. $P_\lambda(\cdot)$ 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 (15) using xtune LASSO, then predict response variable class by (16).

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 $\operatorname{Var}(X\beta)/\sigma^2$.
2. The informativeness of the external metadata, which is controlled by the number of non-zero hyperparameters α .
3. The number of predictor features p .
4. The proportion of predictive features, captured by the number of non-zero true regression coefficients β .
5. The overall magnitude of the non-zero entries of the true regression coefficient vector β , which is controlled by α_0 .
6. The overall degree of correlation between features.

The simulation data was generated according to the following steps: 1) set the parameter α controlling the informativeness of the external metadata; 2) generate the external metadata matrix Z ; 3) generate the vector of regression coefficients β based on α and Z ; 4) generate the feature data X ; 5) generate the outcome Y based on β , X and an independent drawn random error.

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 j^{th} 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 was appended to Z corresponding to the 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 $b = \exp(Z_j \alpha)$. Threshold as described in Reid *et al.* (2013) was then applied to achieve sparsity. Specifically, the $[n^\delta]$ largest β s in magnitude were retained, while all smaller entries were set to 0. The parameter δ controls the sparsity of the final β vector, with a lower value of δ inducing a sparser vector. 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 $Y = X\beta + \epsilon$. Each simulated data is split into a training set ($n = 200$) and a test set ($n = 1000$). The performance of the xtune LASSO was assessed 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, which was fitted to compare with the proposed xtune LASSO, was performed by 10-fold cross-validation. One hundred replicates were generated for each scenario.

We consider nine scenarios varying each of the main simulation parameters described above:

1. $SNR = 1, 3, 5$
2. $p = 500, 1000, 3000$
3. External data informativeness: low, medium, and high. We fixed the sub-set of the $\alpha = (\alpha_0, -1, -0.78, -0.56, -0.33, -0.11, 0.11, 0.33, 0.56, 0.78, 1)$. The low external information is simulation from $\alpha_{low} = (\alpha, \underbrace{0, \dots, 0}_{40})$. The medium external information is simulated from $\alpha_{medium} = (\alpha, \underbrace{0, \dots, 0}_{20})$, and the high external information is simulated from $\alpha_{high} = \alpha$. 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 $\rho = 0, 0.3, 0.6, 0.9$

The simulation results are summarized in Figure 1.

3.1.2 Simulation results and data application of EB tuning

Figure 1 (a) shows box plots (across simulation replicates) of the prediction R^2 for the standard and xtune LASSO as a function of the SNR of the metadata. As the SNR the prediction performance increases for both methods but the xtune LASSO has a better prediction across all levels of SNR considered. When the p/n ratio gets higher, both methods have decreased performance. However, xtune LASSO has a

Figure 1: Simulation Results. Subplot (a): SNR = 1, 3, 5, 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 = 3, $q = 10$, $\alpha_0 = 3$, $\delta = 0.5$, $\rho = 0.2$. Subplot (c): $q = 10, 30, 50$, with $n = 200$, $p = 1000$, SNR = 3, $\alpha_0 = 3$, $\delta = 0.5$, $\rho = 0.2$. Subplot (d): regression coefficients sparsity $\delta = 0.3, 0.5, 0.7$, with $n = 200$, $p = 1000$, SNR = 3, $q = 10$, $\alpha_0 = 3$, $\rho = 0.2$. Subplot (e): overall penalty magnitude $\alpha_0 = 1, 3, 5$, $n = 200$, $p = 1000$, SNR = 3, $q = 10$, $\delta = 0.5$, $\rho = 0.2$. Subplot (f): $\rho = 0, 0.3, 0.6, 0.9$, with $n = 200$, $p = 1000$, SNR = 3, $q = 10$, $\alpha_0 = 3$, $\delta = 0.5$.

slower decreasing rate. When the p/n ratio is very high, we see that the performance of standard LASSO decreased dramatically and xtune LASSO significantly better prediction performance than standard LASSO. The R^2 for standard LASSO remain the same across different level of external data informativeness. The performance of xtune LASSO decreases when the external information has low informativeness. In the case when the external information is completely useless, xtune LASSO performs slightly worse than standard LASSO (not shown).

Figure 1 (d) shows the effect of decreased β sparsity. The higher 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 $\delta = 0.5$, 14 out of 1000 features are non-zero. Both methods have a decreased performance when the model has more signal features, but xtune LASSO has a slower decreasing rate. α_0 controls the overall amount of shrinkage. From Figure 1 (e), the overall amount of shrinkage seems to have little effect on the performance of both methods. Notice that both methods have a higher prediction ability when the signal features (non-zero) features are correlated with each other. This is especially true for standard LASSO. The improved prediction of xtune LASSO over standard LASSO becomes smaller as the correlation between signal features ρ_{signal} 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) studies the influence of correlation on LASSO prediction, and they suggest that correlation in features is not problematic for LASSO prediction. They also find that the prediction errors are mostly smaller for the correlated settings in experimental studies.

In summary, our simulation results show that the LASSO with individual penalties informed by meta-features can outperform the standard LASSO in terms of prediction when 1) the meta-features are informative for the regression effect sizes, 2) the true model is less sparse, and 3) 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. The bone density is measured by the hip T-score derived from the women’s biopsies, with a higher score indicating higher bone density. The microarray data contains over 22,815 probe sets. The data 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.

Insights from previous study results are used as external covariates. (Reppe *et al.*, 2010) identified eight genes that are highly associated with bone density variation. We create a two-column external information Z . The first column is a column of

all 1, representing the overall amount of shrinkage. The second column is a binary variable indicating whether each gene is one of the eight genes identified.

To illustrate the advantage of incorporating external information, we compare our proposed method to the standard LASSO and also the adaptive LASSO. The adaptive LASSO also adjusts the penalty individually for each coefficient, but the adaptive weight is obtained from the initial estimate of the coefficients using the same data rather than external information. Our implementation of adaptive LASSO utilizes the *adalasso()* function in the *parcor* R package.

The data were randomly split into a training data consisting of 80% of the observations and a test data set 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.

Figure 2: Compare prediction MSE, R^2 and number of selected covariates of adaptive LASSO, standard LASSO and xtune LASSO using bone density data. The median R^2 is 0.27 for adaptive LASSO; 0.38 for the standard LASSO, and 0.43 for xtune LASSO. The median number of selected covariates is 10 for adaptive LASSO, 43 for standard LASSO and 12 for the xtune LASSO.

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.

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 is 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 phenotypic 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. Therefore, the meta-feature matrix $\mathbf{Z}$ consists of four external covariates: the first covariates is a binary variable indicating whether a predictor in $\mathbf{X}$ is a clinical feature (1) or a gene expression feature (0) and the next three columns map genes that belong to with 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.

Figure 3: ROC curves for adaptive LASSO, standard LASSO and xtune LASSO applied on the breast cancer dataset

Table 1 compares the AUC, the number of selected features and the computa-

tion 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 three clinical variables are given a very small penalty in xtune LASSO and are both estimated to be non-zero. The penalty chosen by repeated cross-validation for the standard LASSO is 0.020. For xtune LASSO, the penalty applied to clinical variables, ‘non-attractor’ genes, LYM metagenes, MES metagenes and CIN metagenes is 0.001, 0.114, 0.030, 0.038 and 0.028, respectively. This illustrates how the xtune LASSO can induce differential shrinkage among the features according to their empirical importance. In this example, 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 10 selected features while the standard LASSO selected 207 features. However, the xtune LASSO is more computationally demanding than the standard LASSO tuned by cross-validation.

	Standard LASSO	Adaptive LASSO	xtune LASSO
AUC	0.675	0.718	0.746
No. of selected covariates	207	5	10
Computation time	1.58	30.81	12.15

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

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 can yield better prediction performance than standard LASSO. These findings are consistent with the work of (van de Wiel *et al.*, 2016) who proposed a method for estimating group-specific penalties for ridge regression and showed that the use of prior knowledge can improve prediction performance.

xtune LASSO differs from related methods (Pan *et al.*, 2010), (Boulesteix *et al.*, 2017), (Liu *et al.*, 2018) 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 1s) xtune performs empirical Bayes tuning of the single LASSO penalty parameter, providing an alternative to standard tuning by cross-validation. However, 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.

In both our real data application examples, the meta-features are 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. However, xtune is not limited to categorical meta-features and can equally handle quantitative ones, such as effect estimates or p-values for the association of the features and the outcome of interest or related outcomes, which can be exploited to integrate information from previous studies.

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 yield sparser and therefore more interpretable models than standard LASSO regression. Moreover, the estimates of the meta-feature coefficients α can yield additional biological insights as they capture the importance of the meta-features for predicting the outcome.

However, we did note that when the number of meta-features q relative to the sample size n is large, the α 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 5000, with p 50,000 features and q 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.

References

- Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. the gene ontology consortium. *Nat Genet*, **25**(1), 25–29.
- Bergersen, L. C., Glad, I. K., and Lyng, H. (2011). Weighted lasso with data integration. *Statistical Applications in Genetics and Molecular Biology*, **10**(1).
- Boulesteix, A.-l., Bin, R. D., Jiang, X., Fuchs, M., Boulesteix, A.-l., Bin, R. D., Jiang, X., and Fuchs, M. (2017). IPF-LASSO : integrative L 1 -penalized regression with penalty factors for prediction based on multi-omics data IPF-LASSO : integrative L 1 -penalized regression with penalty factors for prediction based on multi-omics data. **2017**(187).
- Boyd, S. and Vandenberghe, L. (2004). *Convex Optimization*. Cambridge University Press, New York, NY, USA.
- Cheng, W., Ou Yang, T.-H., and Anastassiou, D. (2013). Biomolecular events in cancer revealed by attractor metagenes. *PLoS computational biology*, **9**, e1002920.
- Curtis, C., Shah, S. P., Chin, S. F., Turashvili, G., Rueda, O. M., Dunning, M. J., Speed, D., Lynch, A. G., Samarajiwa, S., Yuan, Y., Gräf, S., Ha, G., Haffari, G., Bashashati, A., Russell, R., McKinney, S., Aparicio, S., Brenton, J. D., Ellis, I., Huntsman, D., Pinder, S., Murphy, L., Bardwell, H., Ding, Z., Jones, L., Liu, B., Papatheodorou, I., Sammut, S. J., Wishart, G., Chia, S., Gelmon, K., Speers, C., Watson, P., Blamey, R., Green, A., MacMillan, D., Rakha, E., Gillett, C., Grigoriadis, A., De Rinaldis, E., Tutt, A., Parisien, M., Troup, S., Chan, D., Fielding, C., Maia, A. T., McGuire, S., Osborne, M., Sayalero, S. M., Spiteri, I., Hadfield, J., Bell, L., Chow, K., Gale, N., Kovalik, M., Ng, Y., Prentice, L., Tavaré, S., Markowitz, F., Langerød, A., Provenzano, E., Purushotham, A., Børresen-Dale, A. L., and Caldas, C. (2012). The genomic and transcriptomic architecture of 2,000 breast tumours reveals novel subgroups. *Nature*, **486**(7403), 346–352.
- Forbes, S., Bindal, N., Bamford, S., Cole, C., Kok, C., Beare, D., Jia, M., Shepherd, R., Leung, K., Menzies, A., Teague, J., Campbell, P., Stratton, M., and Futreal, P. (2010). Cosmic: Mining complete cancer genomes in the catalogue of somatic mutations in cancer. *Nucleic acids research*, **39**, D945–50.
- Foster, S. D., Verbyla, A. P., and Pitchford, W. S. (2008). A random model approach for the LASSO. *Computational Statistics*, **23**(2), 217–233.
- Hebiri, M. and Lederer, J. (2013). How correlations influence lasso prediction. *IEEE Transactions on Information Theory*, **59**(3), 1846–1854.
- Jerome Friedman, Trevor Hastie and Tibshirani, R. (2010). Regularization Paths for Generalized Linear Models via Coordinate Descent. **33**(11), 1212–1217.
- 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 Bioinformatics*, **15**.

- Le Thi, H. A. and Pham Dinh, T. (2018). Dc programming and dca: thirty years of developments. *Mathematical Programming*, **169**.
- Li, Q. and Lin, N. (2010). The bayesian elastic net. *Bayesian Analysis*, **5**.
- Liu, J., Liang, G., Siegmund, K. D., and Lewinger, J. P. (2018). Data integration by multi-tuning parameter elastic net regression. *BMC Bioinformatics*, **19**(1), 369.
- MacKay, D. J. C. (1992). Bayesian Interpolation. *Neural Computation*, **4**(3), 415–447.
- Mai, Q., Zou, H., and Yuan, M. (2012). A direct approach to sparse discriminant analysis in ultra-high dimensions. *Biometrika*, **99**(1), 29–42.
- Neal, R. M. (1995). by Bayesian Learning for Neural Networks.
- Nuyten, D., Kreike, b., Hart, A., Chi, J.-T., Sneddon, J., Wessels, L., Peterse, H., Bartelink, H., Chang, H., and Vijver, M. (2006). Predicting a local recurrence after breast-conserving therapy by gene expression profiling. breast cancer res 8:r62. *Breast cancer research : BCR*, **8**, R62.
- Pan, W., Xie, B., and Shen, X. (2010). Incorporating predictor network in penalized regression with application to microarray data. *Biometrics*, **66**(2), 474–484.
- Park, T. and Casella, G. (2008). The Bayesian Lasso. *Journal of the American Statistical Association*, **103**(482), 681–686.
- Pletscher-Frankild, S., Paljeja, A., Tsafou, K., Binder, J., and Jensen, L. (2014). Diseases: Text mining and data integration of disease–gene associations. *Methods (San Diego, Calif.)*, **74**.
- Reid, S., Tibshirani, R., and Friedman, J. (2013). A Study of Error Variance Estimation in Lasso Regression. pages 1–30.
- Reppe, S., Refvem, H., Gautvik, V. T., Olstad, O. K., Høvring, P. I., Reinholt, F. P., Holden, M., Frigessi, A., Jemtland, R., and Gautvik, K. M. (2010). Eight genes are highly associated with BMD variation in postmenopausal Caucasian women. *Bone*, **46**(3), 604–612.
- Rouillard, A., Gundersen, G., Fernandez, N., Wang, Z., Monteiro, C., McDermott, M., and Ma’ayan, A. (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**(14), 1775–1782.
- Tharmaratnam, K., Sperrin, M., Jaki, T., Reppe, S., and Frigessi, A. (2016). Tilting the lasso by knowledge-based post-processing. *BMC Bioinformatics*, **17**(1), 1–9.
- Tibshirani, R. (1996). Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **58**(1), 267–288.
- Tipping, M. (2001). Sparse Bayesian Learning and the Relevance Vector Mach. *Journal of Machine Learning Research*, **1**, 211–244.

- van de Wiel, M. A., Lien, T. G., Verlaet, W., van Wieringen, W. N., and Wilting, S. M. (2016). Better prediction by use of co-data: Adaptive group-regularized ridge regression. *Statistics in Medicine*, **35**(3), 368–381.
- Wipf, D. and Nagarajan, S. (2008). A New View of Automatic Relevance Determination. *Compute*, **20**(2), 1625–1632.
- Wipf, D. and Nagarajan, S. (2014). Iterative Reweighted and Methods for Finding Sparse Solutions Iterative Reweighted Methods for Finding Sparse Solutions. (August), 1–29.
- Yuan, M. and Lin, Y. (2006). Model selection and estimation in regression with grouped variables. *Journal of the Royal Statistical Society. Series B: Statistical Methodology*, **68**(1), 49–67.
- Zeng, C. and Lewinger, J. P. (2019). *xtune: Regularized Regression with Differential Penalties Integrating External Information*. R package version 1.0.0.
- Zou, H. (2006). The adaptive lasso and its oracle properties. *Journal of the American Statistical Association*, **101**(476), 1418–1429.