xrnet: Hierarchical Regularized Regression to Incorporate External Data

Garrett M Weaver[1] **and Juan Pablo Lewinger**[1]

1 Department of Preventive Medicine, University of Southern California

DOI: 10.21105/joss.01761

Software

- Review
- Repository
- Archive

Editor: Melissa Gymrek

Reviewers:

- @juanvillada
- @jsgalan

Submitted: 08 September 2019
Published: 06 December 2019

Summary

Regularized regression is an essential tool for both feature selection and prediction with high-dimensional data. A number of R (R Core Team, 2019) packages have been developed to fit regularized regression models, including `glmnet` (Friedman, Hastie, & Tibshirani, 2010), `biglasso` (Zeng & Breheny, 2017), and `ncvreg` (Breheny & Huang, 2011). These packages can fit multiple model types including linear, multivariate linear, logistic, and Cox regression with different regularization penalties. The penalties control the complexity of the models by shrinking the coefficients toward zero, with the degree of shrinkage controlled by a tuning parameter that is typically selected by cross-validation. In addition to shrinkage, penalties like the lasso, elastic-net, SCAD (Fan & Li, 2001), and MCP (Zhang, 2010) also perform feature selection by shrinking some coefficients to exactly zero.

In statistical genetics and bioinformatics, there has been an increased interest in extending regularized regression methods to integrate external data that may be informative for the association of high-dimensional genomic features (i.e., gene expression, methylation, genotypes) with a health-related outcome (i.e., cancer recurrence). Potential sources of external information within these domains include genomic annotations that describe the underlying functions of a genomic region, and summary statistics derived from external data sources or previous studies. The primary interest is to exploit the external data to both improve the estimation of the regression coefficients and increase the overall predictive performance of the fitted model.

The `xrnet` R package implements a novel extension of regularized regression that enables the integration of meta-features, a particular type of external data. Meta-features, also known as meta-variables or co-data, refer to characteristics of the predictor variables. For example, meta-features of a gene expression variable can be the known function/s of the particular gene. Meta-features of a single nucleotide polymorphism (SNP) genotype variable could be information about whether the SNP is within a regulatory region. Let y be an n-dimensional outcome vector, X be a set of p potential predictors measured for the n observations, and Z be a set of q meta-features available for the p predictors. Our model is related to a standard two-level hierarchical regression model, where the mean effects of the predictors, β, on a outcome are assumed to be dependent on the set of meta-features, Z, through a second set of regression coefficients, α.

$$y = X\beta + \epsilon$$

$$\beta = \alpha_0 1_p + Z\alpha + \gamma$$

where ϵ and γ are error terms and 1_p denotes a vector of ones of dimension p. As a concrete example, assume that X is a set of gene expression features measured on n subjects and that Z is a single meta-feature, Z_1, consisting of a 0-1 indicator variable for whether each gene in X has function “A”. In this simple case, the hierarchical model assumes that the mean effect of expression on the outcome among genes with function "A" is $\alpha_0 + \alpha_1$ and the effect among genes that do not have function "A" is α_0.

The general form of our model extends this two level hierarchy to a high-dimensional setting by jointly modeling X and Z in a regularized regression framework that accounts for the hierarchical nature of the data. In the case of a continuous outcome, the model can be expressed by the following convex optimization problem.

$$\min_{\beta,\alpha_0,\alpha} \frac{1}{2}||y - X\beta||_2^2 + \frac{\lambda_1}{r}\sum_{j=1}^{p}|\beta_j - \alpha_0 - Z_j^T\alpha|^r + \frac{\lambda_2}{s}\sum_{k=1}^{q}|\alpha_k|^s$$

In the joint minimization above, λ_1 and λ_2 are hyperparameters and $r, s = 1(lasso), 2(ridge)$ determine the type of regularization at each level. `xrnet` can also penalize either level with an elastic-net penalty as well. Unlike standard regularized regression, the predictor coefficients, β, are not shrunk towards zero, but rather towards $\alpha_0 1_p + Z\alpha$ as λ_1 increases. The third term in the model allows for variable selection of the 'meta-features' and can shrink α towards zero. To efficiently solve this convex optimization problem for various hyperparameter combinations, the variable substitution $\gamma = \beta - \alpha_0 I_p - Z\alpha$ is used to re-express the problem. The objective function is then a standard regularized regression where the type of regularization and hyperparameter values are variable-specific.

$$\min_{\gamma,\alpha_0,\alpha} \frac{1}{2}||y - X\gamma - \alpha_0 X 1_p + XZ\alpha)||_2^2 + \frac{\lambda_1}{r}\sum_{j=1}^{p}|\gamma_j|^r + \frac{\lambda_2}{s}\sum_{k=1}^{q}|\alpha_k|^s$$

This package extends the coordinate descent algorithm of Friedman (Friedman et al., 2010) to allow for this variable-specific penalization in order to fit the model described above.

Along with this extension, `xrnet` can fit standard regularized regression models and integrates popular features from the R packages `glmnet` and `biglasso`. Below is a comparison of features that are available in `xrnet`, `glmnet`, and `biglasso`. In addition to continuous and binary outcomes, there is active development to extend `xrnet` to survival outcomes, including Cox regression and accelerated failure time models.

Feature	xrnet	glmnet	biglasso
Matrix types supported	Dense (In-Memory), Sparse (In-Memory), Memory-mapped	Dense (In-Memory), Sparse (In-Memory)	Memory-mapped
Outcome types supported	Gaussian, Binomial	Gaussian, Multiresponse Gaussian, Binomial, Poisson, Cox	Gaussian, Binomial
Feature-specific penalty scaling	yes	yes	yes
Feature-specific penalty types	yes	no	no
User controls feature standardization	yes	yes	no
User controls inclusion of intercept	yes	yes	no

Feature	xrnet	glmnet	biglasso
Box (upper/lower constrains on) estimates	yes	yes	no
Enhanced feature screening	no	no	yes
Integration of external data	yes	no	no

The core functionality of the package is written in C++ with integration to R by using the Rcpp R package (Eddelbuettel & François, 2011). The Eigen linear algebra library (Guennebaud, Jacob, & others, 2010) and RcppEigen (Bates & Eddelbuettel, 2013) R package are utilized to handle the dense and sparse data structures. Overall, this R package aims to provide a set of functions to fit and tune hierarchical regularized regression models and unifies some of the best features from currently available R packages for regularized regression into a single easy to use interface.

Funding and Support

This work is supported by the National Institute of Health (NIH) Grant P01CA196569.

References

Bates, D., & Eddelbuettel, D. (2013). Fast and elegant numerical linear algebra using the RcppEigen package. *Journal of Statistical Software*, *52*(5), 1–24. doi:10.18637/jss.v052.i05

Breheny, P., & Huang, J. (2011). Coordinate descent algorithms for nonconvex penalized regression, with applications to biological feature selection. *Annals of Applied Statistics*, *5*(1), 232–253. doi:doi:10.1214/10-AOAS388

Eddelbuettel, D., & François, R. (2011). Rcpp: Seamless R and C++ integration. *Journal of Statistical Software*, *40*(8), 1–18. doi:10.18637/jss.v040.i08

Fan, J., & Li, R. (2001). Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, *96*(456), 1348–1360. doi:10.1198/016214501753382273

Friedman, J., Hastie, T., & Tibshirani, R. (2010). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. doi:10.18637/jss.v033.i01

Guennebaud, G., Jacob, B., & others. (2010). Eigen v3. http://eigen.tuxfamily.org.

R Core Team. (2019). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from https://www.R-project.org/

Zeng, Y., & Breheny, P. (2017). The biglasso package: A memory- and computation-efficient solver for lasso model fitting with big data in R. *ArXiv e-prints*. Retrieved from https://arxiv.org/abs/1701.05936

Zhang, C.-H. (2010). Nearly unbiased variable selection under minimax concave penalty. *The Annals of Statistics*, *38*(2), 894–942. doi:10.1214/09-AOS729