Bioinformatics Advance Access published April 4, 2016

Bioinformatics, 2016, 1–3
doi: 10.1093/bioinformatics/btw143
Advance Access Publication Date: 15 March 2016
Applications Note

Genetics and population analysis

EBglmnet: a comprehensive R package for sparse generalized linear regression models

Anhui Huang* and Dianting Liu

Department of Electrical and Computer Engineering, University of Miami, Coral Gables, FL 33146, USA

*To whom correspondence should be addressed

Associate Editor: Oliver Stegle

Received on September 25, 2015; revised on February 6, 2016; accepted on March 8, 2016

Abstract

Summary: EBglmnet is an R package implementing empirical Bayesian method with both lasso (EBlasso) and elastic net (EBEN) priors for generalized linear models. In our previous studies, both EBlasso and EBEN outperformed other state-of-the-art methods such as lasso and elastic net in inferring sparse genotype and phenotype associations, in which the number of covariates is typically much larger than the sample size. While high density genetic markers can be easily obtained nowadays in genetics and population analysis thanks to the advancements in molecular high throughput technologies, EBglmnet will be a very useful tool for statistical modeling in this area.

Availability and implementation: EBglmnet package is freely available from the R archive CRAN (http://cran.r-project.org/).

Contact: a.huang1@umiami.edu

1 Introduction

In areas such as multiple quantitative trait locus (QTL) mapping and pathway-based genome-wide association studies (GWAS), high density gnomic markers can be obtained easily as technologies advanced in molecular genotyping. Including all marker effects in a single regression model leads to a large number of model variables, typically much larger than the sample size. Traditional regression methods failed when $p \gg n$, where p is the number of model variables, and n is the sample size.

We recently developed an efficient empirical Bayesian method with lasso (EBlasso) (Cai *et al.*, 2011; Huang, 2014; Huang *et al.*, 2013) and elastic net (EBEN) priors (Huang *et al.*, 2015) for variable selection and effect estimation. EBlasso has two sets of hierarchical priors (EBlasso-NE: normal and exponential hierarchical prior; EBlasso-NEG: normal, exponential and gamma hierarchical prior) enforcing different levels of shrinkage to produce a sparse model; and EBEN encourages a grouping effect while inferring a sparse model, which is able to select a group of highly correlated variables when EBlasso fails. Comprehensive simulation and real data analysis in QTL mapping (Cai *et al.*, 2011; Huang *et al.*, 2013, 2014a,b, 2015) and pathway-based GWAS (Huang *et al.*, 2014a) demonstrated that EBlasso and EBEN outperformed other state-of-the-art algorithms such as lasso (Tibshirani, 1996) and elastic net (Zou and Hastie, 2005) in term of power of detecting true effects (PD), false discovery rate (FDR), as well as detecting a group of highly correlated effects.

The aforementioned studies implemented those methods separately either in MATLAB or R, and were not available for different generalized linear regression models (GLMs). In this paper, we developed a cross-platform software package *EBglmnet*, with implementation of both Gaussian and binomial models for all methods. Core algorithms in the software are programmed in efficient C/C++ with user-friendly R interface. Comparing to the original implementation, computational speed via EBglmnet is very fast thanks to the basic linear algebra subprograms (BLAS) and linear algebra package (LAPACK) (Anderson, 1999) utilized by the package.

2 Methods

Let us consider using a GLM for variable selection,

$$\boldsymbol{\eta} = \mu \boldsymbol{I} + \mathbf{X}\boldsymbol{\beta}, \tag{1}$$

where $\mathbf{X}$ an $n \times p$ matrix containing p variables for n samples with p being allowed to be $\gg n$, $\boldsymbol{\eta}$ is an $n \times 1$ linear predictor and is related to an $n \times 1$ outcome variable y through a link function

g: $\mathrm{E}(\boldsymbol{y}|\mathbf{X}) = \mathrm{g}^{-1}(\boldsymbol{\eta}) = \mathrm{g}^{-1}(\mu\boldsymbol{I} + \mathbf{X}\boldsymbol{\beta})$, and $\boldsymbol{\beta}$ is a $p \times 1$ vector capturing the corresponding effects. Based on the model assumptions, the distribution of $\boldsymbol{y}$ depends on the linear predictor $\mu\boldsymbol{I} + \mathbf{X}\boldsymbol{\beta}$, and in some cases, a dispersion parameter $\boldsymbol{\varepsilon}$. For example, in our previous studies of multiple QTL mapping, where $\boldsymbol{y}$ is the quantitative traits, (1) corresponds to the linear regression model:

$$\boldsymbol{y} = \mu\boldsymbol{I} + \mathbf{X}\boldsymbol{\beta} + \boldsymbol{\varepsilon} \tag{2}$$

and $\boldsymbol{\varepsilon}$ can be modelled by an independent identical Gaussian distribution with zero-mean and covariance $\sigma_0^2\mathbf{I}$; In the example of pathway-based GWAS where $\boldsymbol{y}$ indicates the binary outcome of disease status for the n samples, $\boldsymbol{y}$ is linked to (1) via a logit function:

$$\mathrm{logit}(p_i) = log[p_i/(1-p_i)] = \eta_i = \mu + \boldsymbol{x}_i\boldsymbol{\beta}, \tag{3}$$

where $p_i = \Pr(y_i = 1)$, and $\eta_i = \mu + \boldsymbol{x}_i\boldsymbol{\beta}$ is the linear predictor for the ith sample. Based on the above GLMs, both EBlasso and EBEN assign 'peak zero and flat tails' prior distributions to $\boldsymbol{\beta}$, which have a large proportion of the probability mass around zero while still allowing a small decay rate at the two tails (Park and Casella, 2008). For example, EBlasso-NE employs a two-level hierarchical prior distribution: at the first level, β_j follows an independent normal distribution with mean zero and unknown variance $\sigma_j^2 : \beta_j \sim N(0, \sigma_j^2), j = 1, ., p$. At the second level, σ_j^2 follows an independent exponential distribution with a common parameter λ: $p(\sigma_j^2) = \lambda\exp(-\lambda\sigma_j^2)$, $j = 1, \ldots, p$. Then the empirical Bayesian approach employs a coordinate ascent method to find $\widehat{\sigma}_j^2$, the estimate of σ_j^2, that maximizes the marginal likelihood function of σ_j^2. In the iterative process, many σ_j^2, or equivalent β_j are shrunk to exact zero, and the posterior mean and covariance for the non-zero coefficients $\widehat{\boldsymbol{\beta}}$ are obtained through maximizing the marginal posterior of $\boldsymbol{\beta}$ with the given $\widehat{\sigma}_j^2$, $j = 1, \ldots, p$. In EBglmnet, λ controls the degree of decay rate at the two tails (degree of shrinkage) and is determined via cross validation (CV). More details of the methods and prior setups for EBgmnet can be found in the package documentation and the references therein.

2.1 EBlasso-NEG method

We previously reported a powerful EBlasso method using normal, exponential and gamma hierarchical prior (NEG) with both multiple linear regression model and logistic regression model, and demonstrated that it outperformed other state-of-the-art algorithms in term of PD and FDR, such as lasso (Tibshirani, 1996), HyperLasso (Hoggart *et al.*, 2008), BhGLM (Yi and Banerjee, 2009), EB method (Xu, 2007), as well as the Relevant Vector Machine (RVM) (Tipping and Faul, 2003). The NEG prior enforces a strong degree of variable shrinkage, and is able to handle a model with millions of effects as demonstrated in our whole-genome QTL mapping using the linear regression model (Huang *et al.*, 2014b) and pathway-based GWAS with the logistic regression model (Huang *et al.*, 2014a). A unique feature for this prior is that when genetic interactions are considered, a group EBlasso method is available to apply a higher degree of shrinkage on the interaction effects. In EBglmnet, a group parameter in the function is available to turn on this feature.

2.2 EBlasso-NE and the EBEN method

Based on the NEG prior, a two level normal and exponential hierarchical prior (NE) is also developed for the logistic regression model in our previous study (Huang *et al.*, 2013). With only first two-level of the hierarchical NEG prior, EBlasso-NE typically selects more non-zero effects, and results in a less sparse model. Though the original algorithm was developed for logistic regression model, we further implements the NE prior with multiple linear regression model in EBglmnet.

While EBlasso typically selects one variable out of a group of highly correlated variables, we recent derived a novel elastic net prior that is able to encourage a strong grouping effect similar to that of elastic net, but under the empirical Bayesian framework (Huang *et al.*, 2015). Similar as lasso and elastic net, EBlasso-NE becomes a special case of EBEN.

Note that the computational speed for EBglmnet is mainly determined by the number of non-zero effects selected by the model. Without the third layer constraint, EBlasso-NE and EBEN select more variables with small effects, and take longer time to finish the computation (Table 1). Additionally, for a true effect that is highly correlated with other group of effects, EBEN tends to select the whole group, resulting in a smaller estimated effect size with larger P-value on the true effect, and longer computational time.

2.3 Epistasis analysis in EBglmnet

All the aforementioned methods include a parameter 'Epis' to turn on the feature of epistasis analysis. For p genetic variables, this will

Table 1. Performance comparison For EBglmnet and glmnet (lasso/elastic net) packages

Model[a]	Method[b]	*simI*		*simII*		*simIII*	
		PD/FDR	Time(s)	PD/FDR	Time(s)	PD/FDR	Time(s)
Binomial	NEG	0.68/0.07	0.94	0.15/0.00	2.52	0.99/0.04	404.25
	group NEG[c]	–	–	0.45/0.10	23.74	0.99/0.05	381.81
	NE	0.60/0.04	1.57	0.25/0.17	169.16	0.95/0.11	1962.66
	EBEN	0.52/0.04	1.67	0.25/0.17	137.87	0.93/0.09	2013.21
	lasso	0.98/0.40	0.10	–	–	–	–
	elastic net	0.99/0.50	0.10	–	–	–	–
n/p'		300/481		300/115 921		1000/115 921	
Epistasis (Epis)		FALSE		TRUE		TRUE	

[a] Performance metrics were calculated by the mean of 100 replicates of the simulation having optimal parameter pre-determined through CV. A t-test P-value ≤ 0.05 cutoff was used when determining PD and FDR for EBglmnet methods. Simulations were based on a simulated F_2 population with $p = 481$ even spaced (5 cM) markers having 20 true QTL effects (10 main and 10 epistatic effects in the case of Epis = TRUE; effect sizes were randomly generated from unif[2,3]), and computation was performed on a computer cluster including computing nodes with 2.6 GHz Xeon CPU running Linux.

[b] Lasso and elastic net were compared using glmnet package implemented with Fortran code. The package failed to analyze *simII* and *simIII* due to the separation problem in logistic regression.

[c] When Epis = TRUE, the scale hyperparameter for interaction terms is different with that of main effects by a factor of $\sqrt{p(p-1)/2}$.

analyze a total of $p'=p(p+1)/2$ number of effects adding all pairwise interactions using memory efficient algorithmic techniques, such that only main effect matrix is saved in memory and interaction effect is generated dynamically when the particular effect is under evaluation. As illustrated in (Huang *et al.*, 2014b), p can include both additive and dominance effects, and with appropriate genetic coding, all main effects, additive × additive, additive × dominance, dominance × additive and dominance × dominance interactions can be included in the model with 'Epis' on. Generally, a larger sample size is required in epistasis analysis given the $p(p$ - $1)/2$ more candidate effects. Such performance differences are illustrated in the simulation example using a larger sample size in *simIII* comparing with that of *simII*, which are both simulated with 10 main effects and 10 epistasis effects out of the total $p'=p(p+1)/2=115\ 921$ possible effects.

3 Performance

EBglmnet not only incorporates powerful algorithms developed in several of our papers, but also has novel algorithms first reported in this Application Note that make the package completely functional. The list of algorithms is provided in Table 1 using binomial model as an example, and each method is implemented with a CV method to decide the shrinkage parameter(s). Thanks to the C/C++ implementation with the fast *BLAS* and *LAPACK*, the software is also time efficient, considering the high complexity of the algorithms.

All three sets of simulation have $p>n$. However, performance of *simII* is worse than *simI* given the same sample size but more than 200 times more variables. The noise accumulation and spurious correlation are alleviated with increased sample size in *simIII*. With the same number of variables as *simII*, *simIII* also demonstrates that a longer computational time is required given that more non-zero true effects are selected into the model. Of note, severe spurious correlation leads to complete separation in logistic regression model, and glmnet was not able to handle the problem due to the correlation-based discarding rules (Tibshirani *et al.*, 2012).

4 Conclusions

EBglmnet is a powerful package for GLMs that performs variable selection and effect estimation. The algorithms outperform other state-of-the-art methods in terms of PD, FDR, as well as handling high correlated effects. EBglmnet package can be very useful for research studies such as multiple QTL mapping, pathway-based GWAS, among other research areas (Huang *et al.*, 2010; Liu *et al.*, 2007, 2013; Qiusha and Mei-Ling, 2015).

Conflict of Interest: none declared.

References

Anderson,E. (1999) *LAPACK Users' Guide*. Society for Industrial and Applied Mathematics, Philadelphia, PA, USA.

Cai,X. *et al.* (2011) Fast empirical Bayesian LASSO for multiple quantitative trait locus mapping. *BMC Bioinform.*, **12**, 211.

Hoggart,C.J. *et al.* (2008) Simultaneous analysis of all SNPs in genome-wide and re-sequencing association studies. *PLoS Genet.*, **4**, e1000130.

Huang,A. (2014) *Sparse Model Learning for Inferring Genotype and Phenotype Associations*. University of Miami, Coral Gables, FL, USA.

Huang,A. *et al.* (2010) Characterization of arsenic-resistant bacteria from the rhizosphere of arsenic hyperaccumulator *Pteris vittata*. *Can. J. Microbiol.*, **56**, 236–246.

Huang,A. *et al.* (2013) Empirical Bayesian LASSO-logistic regression for multiple binary trait locus mapping. *BMC Genet.*, **14**, 5.

Huang,A. *et al.* (2014a) Detecting genetic interactions in pathway-based genome-wide association studies. *Genet. Epidemiol.*, **38**, 300–309.

Huang,A. *et al.* (2014b) Whole-genome quantitative trait locus mapping reveals major role of epistasis on yield of rice. *PLoS ONE*, **9**, e87330.

Huang,A. *et al.* (2015) Empirical Bayesian elastic net for multiple quantitative trait locus mapping. *Heredity*, **114**, 107–115.

Liu,D. *et al.* (2007) Face Recognition Using Hierarchical Isomap. *2007 IEEE Workshop on Automatic Identification Advanced Technologies*. IEEE, pp. 103–106.

Liu,D. *et al.* (2013) Spatial-temporal motion information integration for action detection and recognition in non-static background. *2013 IEEE 14th International Conference on Information Reuse and Integration (IRI)*. IEEE, pp. 626–633.

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

Qiusha,Z. and Mei-Ling,S. (2015) Sparse linear integration of content and context modalities for semantic concept retrieval. *IEEE Trans. Emerg. Topics Comput.*, **3**, 152–160.

Tibshirani,R. (1996) Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **58**, 267–288.

Tibshirani,R. *et al.* (2012) Strong rules for discarding predictors in lasso-type problems, *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **74**, 245–266.

Tipping,M.E. and Faul,A.C. (2003) Fast marginal likelihood maximisation for sparse Bayesian models. In: Bishop,C.M. and Frey,B.J. (Eds.) *Proceedings of the Ninth International Workshop on Artificial Intelligence and Statistics*. Key West, FL.

Xu,S. (2007) An empirical Bayes method for estimating epistatic effects of quantitative trait loci. *Biometrics*, **63**, 513–521.

Yi,N. and Banerjee,S. (2009) Hierachical generalized linear models for multiple quantitative trait locus mapping. *Genetics*, **181**, 1101–1133.

Zou,H. and Hastie,T. (2005) Regularization and variable selection via the elastic net. *J. R. Stat. Soc. Ser. B: Stat. Methodol.*, **67**, 301–320.