Bioinformatics, 38(12), 2022, 3306–3309
https://doi.org/10.1093/bioinformatics/btac328
Advance Access Publication Date: 16 May 2022
Applications Note

Genetics and population analysis

An R package VIGoR for joint estimation of multiple linear learners with variational Bayesian inference

Akio Onogi [1,*] **and Aisaku Arakawa**[2]

[1]Department of Plant Life Science, Faculty of Agriculture, Ryukoku University, Otsu, Shiga 520-2194, Japan and [2]Division of Animal Breeding and Reproduction Research, Institute of Livestock and Grassland Science, National Agriculture and Food Research Organization, Tsukuba, Ibaraki 305-0901, Japan

*To whom correspondence should be addressed.

Associate Editor: Russell Schwartz

Received on March 17, 2022; revised on April 30, 2022; editorial decision on May 6, 2022; accepted on May 11, 2022

Abstract

Summary: An R package that can implement multiple linear learners, including penalized regression and regression with spike and slab priors, in a single model has been developed. Solutions are obtained with fast minorize-maximization algorithms in the framework of variational Bayesian inference. This package helps to incorporate multimodal and high-dimensional explanatory variables in a single regression model.

Availability and implementation: The R package VIGoR (Variational Bayesian Inference for Genome-wide Regression) is available at the Comprehensive R Archive Network (CRAN) (https://cran.r-project.org/) and at GitHub (https://github.com/Onogi/VIGoR).

Contact: onogiakio@gmail.com

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

In current biology, data of multiple omics can be available from many samples (Hasin *et al.*, 2017; Kim and Tagkopoulos, 2018), and data are also available from environments in fine scales (Gupta and Quan, 2018; Munandar *et al.*, 2017). Thus, it is often of interest to discover important variables from such high-dimensional and multimodal data (Wang *et al.*, 2017) or to use them to predict phenotypes (Riedelsheimer *et al.*, 2012). To model such multimodal and high-dimensional variables in realistic time, flexible and fast regression tools are required. High-dimensional variables can be included in regression models using penalties. Although various penalized regressions have been proposed (e.g. Tibshirani, 1996; Zou and Hastie, 2005), implementation of these methods usually assumes unimodal explanatory variables (e.g. glmnet; Friedman *et al.*, 2010). An extension to multiple learners is feasible in a Bayesian framework because each learner can be a module. Indeed, a popular R package BGLR (Pérez and de los Campos, 2014) allows modeling multimodal explanatory variables using different Bayesian regression methods. However, calculation time will be an issue because the package depends on Markov chain Monte Carlo (MCMC) algorithms.

Here, we provide an R package VIGoR (Variational Bayesian Inference for Genome-wide Regression) which can incorporate multimodal explanatory variables using different regression methods. Solutions are obtained with variational inference which is more time-efficient than MCMC. The package was initially developed to provide variational Bayesian inference for linear regressions (Onogi and Iwata, 2016) and has been updated to incorporate multiple learners. The updates are summarized in Supplementary Methods. VIGoR implements multiple regression methods including Bayesian lasso (BL) (Park and Casella, 2008), extended Bayesian lasso (EBL) (Mutshinda and Sillanpää, 2010), Bayesian Alphabets (BayesA, BayesB and BayesC), Bayesian ridge regression (BRR) and best linear unbiased prediction (BLUP). Bayesian Alphabets will be jargons in quantitative genetics; BayesB and BayesC are regressions with spike and slab priors (Habier *et al.*, 2011; Meuwissen *et al.*, 2001) and BayesA uses *t*-distributions as prior distributions of regression coefficients (Meuwissen *et al.*, 2001). These regression methods were selected because they show different properties. BayesB, BayesC and EBL are suitable for variable selection, and BRR and BLUP are suitable for issues where many variables are involved in the response variable. BL and BayesA tend to show the intermediate properties. See Supplementary Methods for the details of these methods. Users can add these learners simultaneously to a single model as

$$y_i = \sum_{m=1}^{M} f_m\left(\mathbf{x}_{m,i},\ \boldsymbol{\theta}_m\right) + \varepsilon_i,$$

where y_i is the response variables for sample i, M is the number of learners in the model, f_m indicates the mth learner, $\mathbf{x}_{m,i}$ is the explanatory variables for learner m and sample i, $\boldsymbol{\theta}_m$ is the parameters of learner m and ε_i is the residual. The theoretical backgrounds and variational Bayesian algorithms are illustrated in the pdf manual of

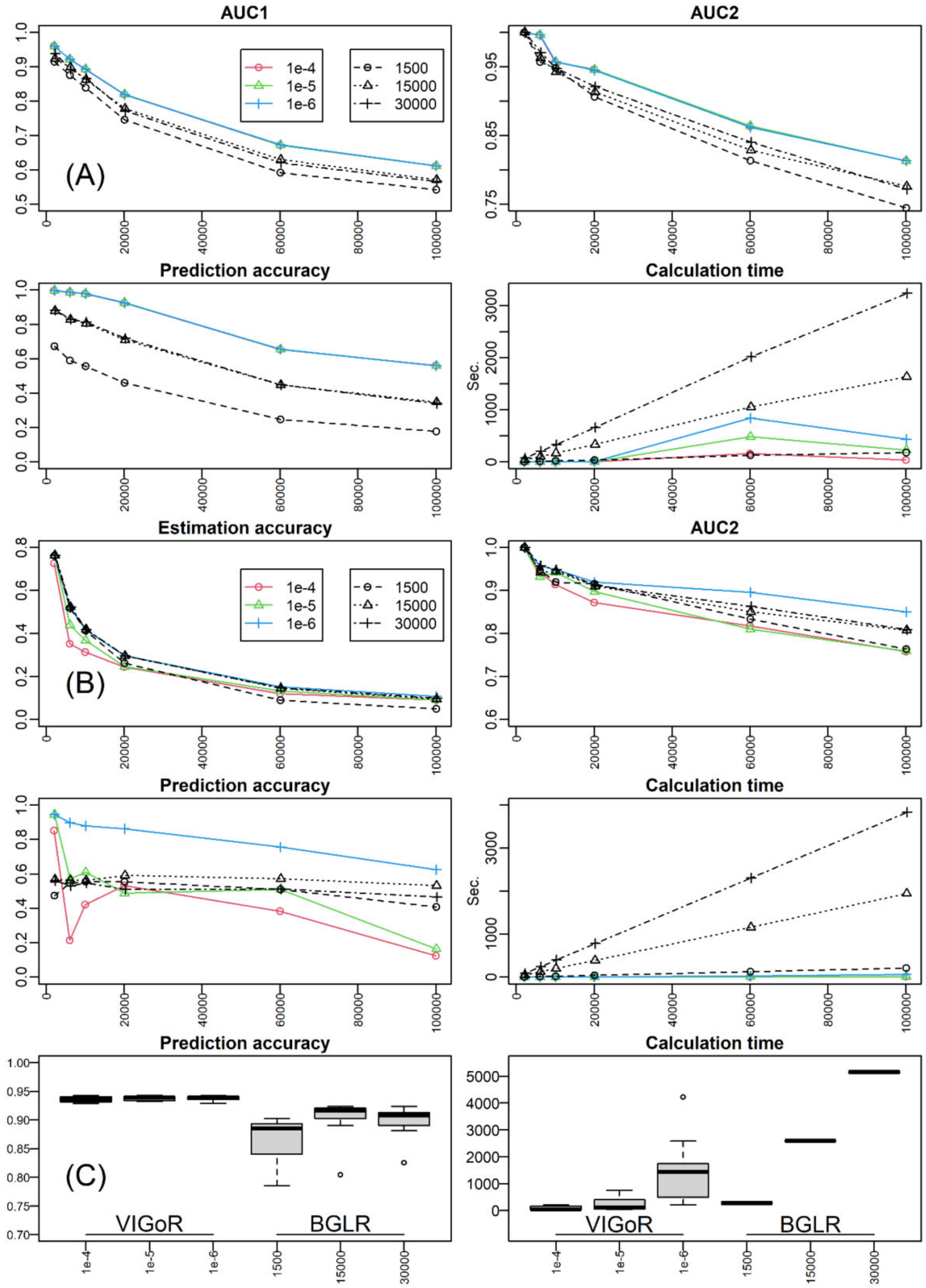

Fig. 1. Experimental results. (**A**) Exp. 1 where BayesC with different shrinkage magnitudes was applied to bimodal explanatory variables. The upper two panels are the AUC for the first and second type variables (AUC1 and AUC2), respectively. The lower two panels are the prediction accuracy and calculation time. For VIGoR and BGLR, three convergence criteria (1e − 4, 1e − 5 and 1e − 6) and chain lengths (1500, 20 000 and 30 000) were attempted, respectively. The *x* axis is the total number of explanatory variables where the bimodal explanatory variables are included half-and-half. Note that AUC1, AUC2 and prediction accuracy of VIGoR were similar among the criteria and thus the curves of 1e − 4 and 1e − 5 are masked by 1e − 6. All plots were averages of 20 replications and the standard deviations are omitted for visual ease (presented in Supplementary Fig. S1). Calculation time was measured with a Windows 10 machine with Intel Core i7-5930K CPU, 3.50 GHz. AUC was calculated using ROCR package (Sing *et al.*, 2005). (**B**) Exp. 2 where Bayesian ridge regression and BayesB were applied to tetra-modal explanatory variables. Estimation accuracy is the Pearson correlation for the first type variable between the true effects and effects estimated by Bayesian ridge regression. AUC2 is the AUC for the second type variable obtained from BayesB. AUC3 and AUC4 (AUC for the third and fourth type variables also obtained from BayesB) are omitted and presented in Supplementary Figure S2. All plots were averages of 20 replications and the standard deviations are presented in Supplementary Figure S2. (**C**) Exp. 3 where BayesC with different shrinkage magnitudes was applied to additive and interaction effects of real soybean data. Prediction accuracy was evaluated using Pearson correlation between the observed and predicted values. The unit of calculation time is second. Runs were repeated 10 times from different initial values

the package available at https://github.com/Onogi/VIGoR. It is notable that the multiple learners in a model can share the explanatory variables (i.e. $\mathbf{x}_{m,i} = \mathbf{x}_{n,i}$ for $m, n \in 1, \ldots, M$ and $m \neq n$). Such modeling would behave like ensemble learning which is expected to be robust against hyperparameter specification and data architecture (Knürr *et al.*, 2013; Onogi *et al.*, 2015).

2 Experiments

Two simulation experiments (Exp. 1 and 2) were performed here. The simulated data in Exp. 1 and 2 were bimodal and tetra-modal, respectively. The total number of explanatory variables was 2000, 6000, 10 000, 20 000, 60 000 or 100 000. Each type (mode) of variables was included in equal numbers and explained the same amount of the variance of the response variable. Both the numbers of training and testing samples were 1000. In Exp. 1, both types of the explanatory variables were generated from the standard normal distribution, and 1% and 0.1% of each variable were assumed to have non-zero effects drawn from the standard normal distribution. BayesC with different shrinkage magnitudes was applied to both types of variables. In Exp. 2, one type of explanatory variable was assumed as SNPs, and genotypes were generated assuming allele frequencies being 0.5. The other three types of variables were generated from the standard normal distribution. All the SNPs were assumed to have non-zero effects, and 0.1% had non-zero effects for the other types of variables. Bayesian ridge regression was applied to SNPs, and BayesB were applied to the other types of variables. BayesB and BayesC implemented in VIGoR and BGLR differ in the inclusion probability of explanatory variables. The probability is fixed to a predefined value in VIGoR whereas it is inferred with a prior beta distribution, Beta(π_0, p_0), in BGLR. Here, π_0 is the expectation and p_0 defines the variance as $\pi_0(1-\pi_0)/(p_0+1)$. For fair comparison, p_0 was set to $1e+6$ to prevent fluctuation of the probability during sampling. The predefined value of VIGoR and π_0 of BGLR were set to the true values of the simulation. VIGoR and BGLR were compared in terms of calculation time, area under the curve (AUC) and prediction accuracy for testing samples. For SNPs in Exp. 2, Pearson correlation between the true and estimated effects was used instead of AUC. These experiments are complementally explained in Supplementary Methods.

Calculation time depends on the criterion of convergence. For VIGoR, three convergence criteria were compared. The iterative update of VIGoR was stopped when $||\boldsymbol{\theta}^* - \boldsymbol{\theta}||^2/||\boldsymbol{\theta}^*||^2 < t$ where $||\,||$ is the Euclidean norm, $\boldsymbol{\theta}$ is the vector containing all parameter values at the previous iteration, $\boldsymbol{\theta}^*$ is the vector consisting of newly updated parameter values at the iteration and t is the convergence criterion which was set to $1e-4$ (loose), $1e-5$ (moderate) or $1e-6$ (strict). Because convergence of MCMC is difficult to verify in particular when many parameters are involved, three chain lengths were compared: (i) nIte = 1500, burnIn = 500 and thin = 5 (default setting of BGLR), (ii) nIte = 15 000, burnIn = 5000 and thin = 10 and (iii) nIte = 30 000, burnIn = 20 000 and thin = 10. Here, nIte, burnIn and thin denote the total number of iterations, length of burnin and sampling interval, respectively.

In Exp. 1, AUC and prediction accuracy of VIGoR were almost same among the convergence criteria (Fig. 1A). AUC and prediction accuracy of BGLR were improved by prolonging the chain length, but still inferior to those of VIGoR despite more calculation time spent. In Exp. 2, VIGoR generally showed the best performance when $t = 1e-6$ (Fig. 1B). Calculation time under this strict criterion was still less than the shortest chain length of BGLR (1500). Complete results with standard deviations are presented in Supplementary Figures S1 and S2.

In Exp. 3, we compared VIGoR with BGLR using a real data of soybean (Onogi *et al.*, 2021). Days to flowering of a major variety (ID V083) evaluated from 2005 to 2015 ($N = 213$) was predicted from records of the variety from 1967 to 2004 ($N = 838$). As explanatory variables, daily mean temperature, photoperiod and precipitation from sowing dates (1st day) to 246th day were used. First-order interactions among these variables were also considered resulting in $246 \times 3 + 246 \times 246 \times 3 = 182\,286$ explanatory variables. These additive and interaction effects were modeled using BayesC with different shrinkage magnitudes. Specifically, 20% and 0.1% of the additive and interaction variables were assumed to have non-zero effects. p_0 of BGLR was set to $1e+6$. The convergence criterion of VIGoR and the chain lengths of BGLR followed those used in Exp. 1 and 2. Prediction accuracy of VIGoR was similar among the convergence criteria (Fig. 1C), and superior to BGLR of any chain length. Calculation time of VIGoR increased as the criterion became stricter, but still less than the longest chain of BGLR that was not able to achieve accuracy equivalent with VIGoR.

3 Conclusions

VIGoR offers fast and accurate solutions for linear models that incorporate multiple learners. The package enables users to model multimodal and high-dimensional explanatory variables, and will help discovering important variables or predicting phenotypes. Although the current version is only applicable to quantitative response variables, future updates will allow application to categorical or censored data.

Funding

This work was supported by Ryukoku University.

Conflict of Interest: none declared.

Data availability

R scripts to reproduce the experiment results and pdf manual of the package are provided at https://github.com/Onogi/VIGoR.

References

Friedman,J. *et al.* (2010) Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.*, **33**, 1–22.

Gupta,G.S. and Quan,V.M. (2018) Multi-sensor integrated system for wireless monitoring of greenhouse environment. In: *2018 IEEE Sensors Applications Symposium (SAS), Seoul*, pp. 1–6.

Habier,D. *et al.* (2011) Extension of the Bayesian alphabet for genomic selection. *BMC Bioinformatics*, **12**, 186.

Hasin,Y. *et al.* (2017) Multi-omics approaches to disease. *Genome Biol.*, **18**, 83.

Kim,M. and Tagkopoulos,I. (2018) Data integration and predictive modeling methods for multi-omics datasets. *Mol. Omics*, **14**, 8–25.

Knürr,T. *et al.* (2013) Impact of prior specifications in a shrinkage-inducing Bayesian model for quantitative trait mapping and genomic prediction. *Genet. Sel. Evol.*, **45**, 1–16.

Meuwissen,T.H. *et al.* (2001) Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, **157**, 1819–1829.

Munandar,A. *et al.* (2017) Design of real-time weather monitoring system based on mobile application using automatic weather station. In: *2017 2nd International Conference on Automation, Cognitive Science, Optics, Micro Electro-Mechanical System, and Information Technology (ICACOMIT), Jakarta*, pp. 44–47.

Mutshinda,C.M. and Sillanpää,M.J. (2010) Extended Bayesian LASSO for multiple quantitative trait loci mapping and unobserved phenotype prediction. *Genetics*, **186**, 1067–1075.

Onogi,A. and Iwata,H. (2016) VIGoR: variational Bayesian inference for genome-wide regression. *J. Open Res. Softw.*, **4**, e11.

Onogi,A. *et al.* (2015) Exploring the areas of applicability of whole-genome prediction methods for Asian rice (*Oryza sativa* L.). *Theor. Appl. Genet.*, **128**, 41–53.

Onogi,A. *et al.* (2021) A method for identifying environmental stimuli and genes responsible for genotype-by-environment interactions from a large-scale multi-environment data set. *Front. Genet.*, **12**, 803636.

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

Pérez,P. and de los Campos,G. (2014) Genome-wide regression and prediction with the BGLR Statistical package. *Genetics*, **198**, 483–495.

Riedelsheimer,C. *et al.* (2012) Genomic and metabolic prediction of complex heterotic traits in hybrid maize. *Nat. Genet.*, **44**, 217–220.

Sing,T. *et al.* (2005) ROCR: visualizing classifier performance in R. *Bioinformatics*, **21**, 3940–3941.

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

Wang,C. *et al.* (2017) Metabolome-wide association study identified the association between a circulating polyunsaturated fatty acid variant rs174548 and lung cancer. *Carcinogenesis*, **38**, 1147–1154.

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