Smooth-threshold multivariate genetic prediction with unbiased model selection

Masao Ueki* and Gen Tamiya†

Abstract — **We develop a new genetic prediction method, smooth-threshold multivariate genetic prediction, using single nucleotide polymorphisms (SNPs) data in genome-wide association studies (GWASs). Our method consists of two stages. At the first stage, unlike the usual discontinuous SNP screening as used in the gene score method, our method continuously screens SNPs based on the output from standard univariate analysis for marginal association of each SNP. At the second stage, the predictive model is built by a generalized ridge regression simultaneously using the screened SNPs with SNP weight determined by the strength of marginal association. Continuous SNP screening by the smooth-thresholding not only makes prediction stable but also leads to a closed form expression of generalized degrees of freedom (GDF). The GDF leads to the Stein's unbiased risk estimation (SURE) which enables data-dependent choice of optimal SNP screening cutoff without using cross-validation. Our method is very rapid because computationally expensive genome-wide scan is required only once in contrast to the penalized regression methods including lasso and elastic net. Simulation studies which mimic real GWAS data with quantitative and binary traits demonstrate that the proposed method outperforms the gene score method and genomic best linear unbiased prediction (GBLUP), and also shows comparable or sometimes improved performance with the lasso and elastic net being known to have good predictive ability but with heavy computational cost. Application to whole-genome sequencing (WGS) data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) exhibits that the proposed method shows higher predictive power than the gene score and GBLUP methods.**

Keyword: *Genetic prediction; marginal association screening; model selection; smooth-thresholding.*

1 Introduction

Genome-wide association study (GWAS) is a popular tool for discovering disease-susceptibility genes using large number of single nucleotide polymorphisms (SNPs) without prior knowledge. Apart from discovery of susceptibility genes, prediction of individual's phenotype from high-dimensional genetic information, termed as a genetic prediction, is an important task for personalized medicine. Currently, researchers are exploring the most effective way of building genetic prediction models (Purcell *et al.* 2009). In this paper, we develop a new statistical approach, smooth-threshold multivariate genetic prediction, for building genetic predictive models with input of large-scale genome-wide SNPs data.

We consider standard multiple regression model but with high-dimensional predictor variables. To be specific, $y = (y_1, \ldots, y_n)^T$ represent response variables of individual's phenotype data modeled by a conditional distribution given predictor variables $X = (X_1, \ldots, X_p)$ observed for n individuals, in which $X_j = (x_{1,j}, \ldots, x_{n,j})^T$ for $j \in M = \{1, \ldots, p\}$. The conditional expectation of y_i given $x_i = (x_{i,1}, \ldots, x_{i,p})$ is assumed to be a linear combination $\eta\{E(y_i|x_i)\} = x_i\beta$, where η is some known monotone function and β is a vector of regression coefficients. In this paper, we consider linear regression with identity map η for quantitative trait such as clinical characteristics, and logistic regression with logit function η for binary trait such as affected/unaffected status. Each X_j is either genotype at a SNP site or other covariate such as sex, age, body mass index (BMI), smoking status, alcohol consumption and principal components for population stratification. Each SNP can take one of three possible genotypes, gg, gG and GG, where g and G denote minor and major alleles at the SNP site, respectively. If X_j represents the observed count of minor allele g at a SNP site, X_j takes a value from $\{0, 1, 2\}$. Under the Hardy–Weinberg equilibrium (HWE), the observed count of minor allele g at each SNP follows a binomial distribution with parameter $f \in [0, 0.5]$ called a minor allele frequency (MAF), i.e. frequency of the minor allele g in general population. Quality controls (QCs) are often conducted to remove low-quality SNPs by checking HWE and missing rates as well as low MAF SNPs. Even after those QCs, large number of SNPs still remain. Since sample sizes are usually far less than the number of SNPs, the predictive

*Biostatistics Center, Kurume University, 67 Asahi-Machi, Kurume, Fukuoka 830-0011, Japan, E-mail: uekimrsd@nifty.com
†Tohoku Medical Megabank Organization, Tohoku University, 2-1 Seiryo-Machi, Aoba-Ku, Sendai 980-8573, Japan

modeling in GWAS faces the $p \gg n$ problem. The $p \gg n$ condition hampers multiple regression that fits simultaneously using p predictors X.

Standard GWAS analysis conducts marginal association scan between y and each X_j independently, i.e. a univariate analysis which tests the slope parameter in univariate regression model, followed by multiple test using a Bonferroni correction with a stringent significance level (e.g. p-value less than 5×10^{-8}) in order to control the rate of false positive findings. Meanwhile, suppose that X does not include covariates and consists of SNPs only. Let $T_j(y, X)$ represent a non-negative test statistic for testing association between jth SNP X_j and y as a function of y and X, and the corresponding inclusion threshold be $t > 0$. For example, t is a chi-squared quantile at a given p-value cutoff for chi-squared test statistics $T_j(y, X)$. The resulting SNP set from a marginal association screening at a threshold t is defined by $A = \{j \in M : T_j(y, X) > t\}$. Purcell *et al.* (2009) proposed a gene score method which simply averages each genotype data weighted by estimated effect size for each SNP in A. Warren *et al.* (2013) consider multiple regression for SNPs in A, called a multivariate gene score method.

In the purpose of prediction, the cutoff t can be chosen in terms of prediction ability. However, evaluating prediction ability is not straightforward unlike in traditional setting without screening. It is known that, the screening invalidates traditional statistical procedures, called an winner's curse effect. Analogous problem arises in the context of prediction modeling. Actually, simulation studies as well as examination on real GWAS datasets reported that screening leads to overfitting. In Ueki and Tamiya (2016), we show that the screening can deflate the residual sum of squares (RSS) compared with the RSS without screening, so that the RSS becomes too optimistic. Since screening complicates the behavior of RSS, naive use of RSS is unwarranted in measuring prediction ability. Instead, we can use cross-validation (or sample splitting) which divides the training data into two parts, one of which is used for ranking SNPs and remaining is used to construct a predictive model. Purcell *et al.* (2009) choose an optimal inclusion cutoff by cross-validation.

Although cross-validation takes into account of the screening, reduced sample sizes in training stage may lose predictive power, which is a severe concern when sample sizes are small. Five or ten-folds cross-validation is commonly used in model selection. For example, the `SparSNP` program (Abraham *et al.*, 2012) implementing penalized regression methods, the lasso and elastic net, searches for entire genome-wide SNPs data without SNP screening. `SparSNP` selects the tuning parameter by k-fold cross-validation with default setting of $k = 10$. Repeated genome-wide scans needed at each candidate tuning parameter and multiple runs of model fitting in each fold increase computational cost. For large-scale data such as the whole-genome sequencing (WGS) data, heavy computational cost critically limits the applicability although penalized methods are known to give better predictive power than the simpler gene score method.

In this paper, we develop a new predictive modeling approach, a smooth-threshold multivariate genetic prediction, which is really applicable to large-scale genome-wide data such as WGS data while preserving high prediction ability. Our method consists of two stages. At the first stage, our method continuously screens SNPs based on the output from standard univariate analysis for marginal association of each SNP. At the second stage, the predictive model is built by a generalized ridge regression simultaneously using the screened SNPs with SNP weight determined by the strength of marginal association reflecting the uncertainty of inclusion. Since the final predictive model is essentially built in multiple regression model as in the sure independence screening, the correlations between predictor variables are accounted for (See also Warren *et al.* (2013)). Marginal association signal is used only for penalizing each regression coefficient. Our method is very rapid because computationally expensive genome-wide scan is required only once in contrast to the penalized methods which need genome-wide scan several times. Our proposal can be seen as a smoothed version of multiple regression after single SNP-GWAS screening of predictor variables at some p-value cutoff, in which the discontinuous process in screening is replaced by a continuous function. The resulting continuity makes the prediction stable in the sense of Breiman (1996). The continuity in SNP screening also leads to a closed form expression of generalized degrees of freedom (GDF; Ye, 1998), and allows an application of Stein's unbiased risk estimation (SURE). While the Mallows' C_p (Mallows, 1973) with the usual degrees of

freedom is no longer unbiased model selection criterion due to the effect of screening, we can readily construct an unbiased C_p-type model selection criterion using the GDF. It allows data-dependent choice of optimal SNP inclusion cutoff without relying on cross-validation. The effect of screening is properly accounted for by the SURE's unbiasedness. Since no cross-validation is needed, computationally expensive genome-wide scan is required only once in ranking SNPs. We also extend to generalized linear models and propose a loglikelihood-based C_p-type model selection criterion. Simulation studies which mimic real SNP-GWAS data for both quantitative and binary traits show that the proposed method gives better performance than gene score and genomic best linear unbiased prediction (GBLUP) and attains a comparable or sometime improved prediction performance with the lasso and elastic net in `SparSNP` program. Application to large-scale WGS data from Alzheimer's Disease Neuroimaging Initiative (ADNI) exhibits that the proposed method gives higher predictive performance than both the gene score and GBLUP methods.

2 Materials and Methods

Here we consider linear multiple regression model, $y = \mu + \epsilon$, where $\mu = E(y|X) = X\beta$, $\epsilon \sim N(0, \sigma^2 I_n)$, X is a p-dimensional design matrix and β is the corresponding p regression coefficients. Since p is much larger than n in typical GWAS data, some dimensionality reduction is required. Sparsity assuming that many components of β are zero would be a realistic assumption. If susceptible SNPs show relatively large marginal signal, marginal association screening effectively reduces the dimensionality. The gene score method (Purcell *et al.*, 2009) and its multivariate generalization (Warren *et al.*, 2013) use upper-ranked SNPs in marginal association, $A = \{j \in M : T_j(y, X) > t\}$, for a given cutoff value $t > 0$. Although dimensionality is effectively reduced, discontinuity in y present in the screening process in A may incur instability of prediction, i.e. small change in data can make large changes in the prediction (Breiman, 1996). To address the discontinuity issue, we use a smooth-thresholding proposed by Ueki (2009). To be specific, we propose to estimate the regression coefficients by

$$\check{\beta} = \begin{pmatrix} \check{\beta}_A \\ \check{\beta}_{A^c} \end{pmatrix} = \begin{pmatrix} \check{G}_A(I_{|A|} - \check{D}_A)X_A^T y \\ 0 \end{pmatrix}, \tag{2.1}$$

where A^c indicates the complement set of A, $\check{G}_A = \{(I_{|A|} - \check{D}_A)(\Sigma_{AA} + \lambda I_{|A|}) + \tau \check{D}_A\}^{-1}$, $\Sigma = X^T X$, $\Sigma_{AA} = (\Sigma_{jk})_{j\in A, k\in A}$, γ and τ are non-negative tuning parameters and $\lambda > 0$ is a small constant to avoid singularity of $\check{G}_A$. The corresponding prediction of y_i is then $\check{\mu}_i(y) = X_i^T \check{\beta}$. Here $\check{D}_j$ is an adaptive lasso smooth-thresholding function defined by

$$\check{D}_j = \min[1, \{t/T_j(y, X)\}^{\frac{1+\gamma}{2}}]. \tag{2.2}$$

Since $\check{D}_j = 1$ if and only if $T_j(y, X) \le t$, the screened set A with $\check{D}_j$ is the same as that with $\hat{D}_j = 1_{\{T_j(y,X)\le t\}}$, where $1_{\{\cdot\}}$ denotes the indicator function. It can be seen that $\check{D}_j$ replaces the discontinuous screening process $\hat{D}_j$ by a continuous function. As a result, $\check{\mu}_i(y)$ turns out to be continuous in y.

The regression coefficient for the screened set in (2.1), $\check{\beta}_A$, can be seen as a solution to

$$X_A^T(X_A\check{\beta}_A - y) + W_A\check{\beta}_A = 0, \tag{2.3}$$

with $W_A = \mathrm{diag}(W_j : j \in A)$ where $W_j = \lambda + \tau\check{D}_j/(1 - \check{D}_j)$, which is the minimizer of a generalized ridge regression loss, $\|y - X_A\beta_A\|^2 + \sum_{j\in A}\beta_j^2 W_j$, with respect to β_A. Ridge weight for each predictor variable, W_j, represents uncertainty of marginal association screening. If the marginal association is very weak, we have $\check{D}_j \approx 1$ and large W_j, then the corresponding regression coefficient is strongly shrunken towards zero. If the marginal association is strong, we have $\check{D}_j \approx 0$ and $W_j \approx \lambda$, then the corresponding regression coefficient is less penalized. From the fact that the winner's curse effect produces larger selection bias for small regression coefficient, it is expected that the above penalization decreases the selection bias.

Predictive power largely depends on the choice of t. It may be done using cross-validation by dividing a dataset into test and training samples (Warren *et al.*, 2013). Cross-validation takes into account sampling variability due to the screening. However, repeated genome-wide scans to obtain the screened set A needed in cross-validation incurs computational burden. It is also concerned that the reduction in training sample sizes decreases the predictive power of the model. Instead of cross-validation, we propose a C_p-type criterion based on SURE using GDF. The continuity of $\check{\mu}_i(y)$ in y leads to a closed-form expression of GDF. In what follows, we consider p-value cutoff α instead of t by a one-to-one transformation $t = F^{-1}(1-\alpha)$, where F^{-1} is a quantile function of the distribution of $T_j(y, X)$ under the null hypothesis of no marginal association such as F or χ^2 distribution. An optimal α is determined by minimizing the C_p-type criterion within a range of α for search, $[\alpha_{\min}, \alpha_{\max}]$. The proposed procedure is outlined at the end of this section. It is noteworthy that the computational intensive genome-wide scan is required only once in the single-SNP association screening at Step 1. Subsequent Steps 2–4 are performed on the reduced set of SNPs whose single-SNP association p-value is less than $\alpha_{\max}$. More details of this section including formulas, derivations, extension to generalized linear models and additional descriptions are given in Ueki and Tamiya (2016).

Outline of algorithm

Step 1. Perform single-SNP association analysis for p SNPs.

Step 2. Screen SNPs whose single-SNP association p-value is less than $\alpha_{\max}$.

Step 3. Fix γ and τ as suggested in main text, and select an optimal α from candidate values in $[\alpha_{\min}, \alpha_{\max}]$ by minimizing the C_p-type criterion: $C(\alpha) = \sum_{i=1}^{n}\{y_i - \check{\mu}_i(\alpha)\}^2 + 2\sigma^2\mathrm{GDF}(\alpha)$. Explicit formulas are given in Ueki and Tamiya (2016).

Step 4. Compute $\check{\beta}$ by (2.1) using the selected α.

3 Acknowledgment

This work was carried out under the ISM General Cooperative Research 1 (2015-ISM-CRP-1013) and was partially supported by a Grant-in-Aid for Young Scientist (B) (25870074) and Grants-in-Aid for Scientific Research (C) (25330049 and 25460403).

References

[1] Abraham G., Kowalczyk A., Zobel J. and Inouye, M. (2012). SparSNP: fast and memory-efficient analysis of all SNPs for phenotype prediction. *BMC Bioinfomatics*, 13:88.

[2] Breiman L. (1996). Heuristics of instability and stabilization in model selection. *Annals of Statistics*, 24:2350–2383.

[3] Mallows C. L. (1973). Some comments on C_p. *Technometrics*, 15:661–675.

[4] Purcell SM., Wray N. R., Stone J. L., Visscher P. M., O'Donovan M. C., Sullivan P. F., Sklar P. and Consortium International Schizophrenia. (2009). Common polygenic variation contributes to risk of schizophrenia and bipolar disorder. *Nature*, 460:748–52.

[5] Ueki M. (2009). A note on automatic variable selection using smooth-threshold estimating equation. *Biometrika*, 96:1005–1011.

[6] Ueki M. and Tamiya G. (2016). Smooth-threshold multivariate genetic prediction with unbiased model selection. *Genetic Epidemiology*, 40:233–243.

[7] Warren H., Casas J. P., Hingorani A., Dudbridge F. and Whittaker J. (2013). Genetic prediction of quantitative lipid traits: comparing shrinkage models to gene scores. *Genetic Epidemiology*, 38:72–83.

[8] Ye J. (1998). On measuring and correcting the effects of data mining and model selection. *Journal of the American Statistical Association*, **93**, 120–131.