# Chapter 10
# Statistical Prediction and Cross Validation

Prediction is the fundamental purpose of statistical analysis in many situations. Let *X* be a set of independent variables (called predictors) and *y* be the response variable. A model with additive error is described by

$$y = f_\theta(X) + e \tag{1}$$

where $f_\theta(X)$ is a function to model the relationship between *X* and *y*, and *e* is the residual error (vector) with an assumed $N(0, I\sigma^2)$ distribution, $\sigma^2$ is the residual error variance. The function $f_\theta(X)$ can be linear or non-linear with a set of parameters denoted by $\theta$. The parameters are often estimated from a set of data collected from an experiment. The assumption of normal distribution is not required in the least squares method. In this chapter, we are going to learn only the linear model,

$$y = X\beta + e \tag{2}$$

where $\beta$ is a set of regression coefficients (including an intercept). The parameter set is $\theta = \{\beta, \sigma^2\}$. The predicted value of the response variable is

$$\hat{y} = f_{\hat{\theta}}(X) = X\hat{\beta} \tag{3}$$

Let us look at the simplest linear regression with just one predictor. The model is

$$y_j = b_0 + X_j b_1 + e \tag{4}$$

The predicted value is

$$\hat{y}_j = b_0 + X_j \hat{b}_1 = \bar{y} + b_1(X_j - \bar{X}) \tag{5}$$

The variance of the predicted value is

$$\begin{aligned} \mathrm{var}(\hat{y}_j) &= \mathrm{var}(\bar{y}) + (X_j - \bar{X})^2 \,\mathrm{var}(b_1) \\ &= \frac{1}{n}\sigma^2 + \frac{(X_j - \bar{X})^2}{\sum (X_j - \bar{X})^2}\sigma^2 \\ &= \left[\frac{1}{n} + \frac{(X_j - \bar{X})^2}{\sum (X_j - \bar{X})^2}\right]\sigma^2 \end{aligned} \tag{6}$$

The derivation of this equation is based on the assumption that $\mathrm{cov}(\bar{y}, \hat{b}_1) = 0$, which is true but the proof of it is very difficult. Later on when we learn matrix algebra, we will use a different formula of the variance, which can be easily proved numerically to be the same as the above variance.

### Prediction and model goodness of fit

Prediction and model goodness of fit are two different concepts. A model fitting the data very well does not mean it has a good prediction. A complex model with many parameters may fit the data perfectly, but it may be poorly in prediction of a "future observation." For example, Figure 1 below shows that noisy (roughly

linear) data is fitted to both linear and polynomial functions. Although the polynomial function is a perfect fit, the linear version can be expected to generalize better. In other words, if the two functions were used to extrapolate the data beyond the fit data, the linear function would make better predictions

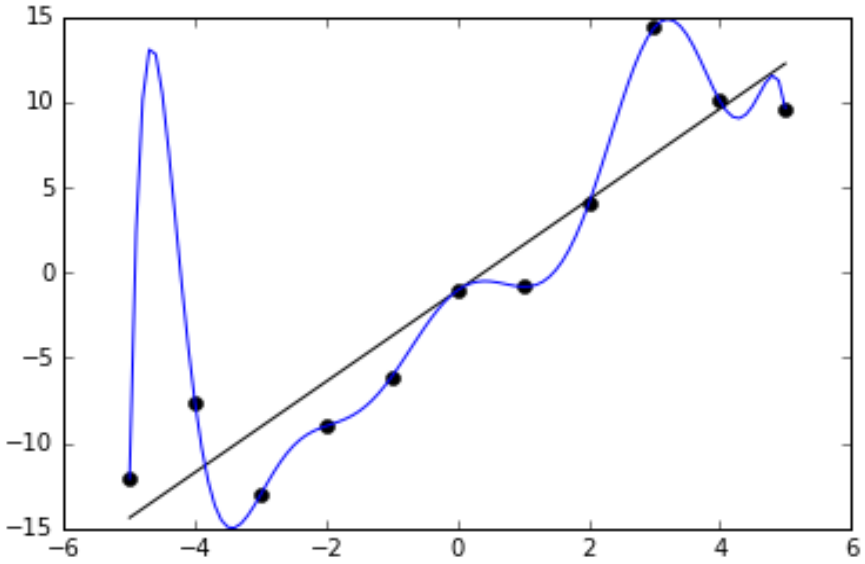


**Figure 1.** Fitting noisy data using a linear model and a high order polynomial model

### Validation and cross validation

To evaluate a model's predictability, it is necessary to have a test sample to validate a model. Imagine that we have a sample to train the model (estimate parameters) and then use the test sample to test the predictability of the model. For example, there are *n* observations in the test sample with *n* observed response data points denoted by $y_j$ for the *j*th data point. Using the parameters from the training sample, we predict the data point denoted by $\hat{y}_j$. The model predictability is measured in one of two ways. The first measurement is the sum of predicted residual errors (PRESS),

$$\mathrm{PRESS} = \sum_{j=1}^{n} (y_j - \hat{y}_j)^2 \tag{7}$$

The smaller the PRESS, the better the model. This measurement depends on the scale of the response variable and the sample size. Therefore, people often convert this measurement into a predictability in a standard scale, such as

$$R^2_{\mathrm{PRESS}} = 1 - \frac{\mathrm{PRESS}}{\mathrm{SS}} \tag{8}$$

where

$$\mathrm{SS} = \sum_{j=1}^{n} (y_j - \bar{y})^2 \tag{9}$$

is the total sum of squares of the response variable. One has to be sure that the average predicted value equals the average of the response, i.e.,

$$\bar{\hat{y}} = \frac{1}{n}\sum_{j=1}^{n} \hat{y}_j = \frac{1}{n}\sum_{j=1}^{n} y_j = \bar{y} \tag{10}$$

If this assumption is violated, you may get an $R^2$ value beyond the range of $(0,1)$.

The second way of measurement is the squared Pearson correlation between the observed and predicted response of the test sample, denoted by

$$R^2_{\mathrm{PEARSON}} = \left\{ \frac{\sum_{j=1}^{n} (y_j - \bar{y})(\hat{y}_j - \bar{\hat{y}})}{\sqrt{\sum_{j=1}^{n} (y_j - \bar{y})^2 \sum_{j=1}^{n} (\hat{y}_j - \bar{\hat{y}})^2}} \right\}^2 \tag{11}$$

In situations where $\bar{y} = \bar{\hat{y}}$, the two measurements are very close to each other. The latter, however, always falls within the range between 0 and 1, and is more preferable than the latter.

The validation approach described here requires a test sample, which is often not available. One may divide the total sample into a training sample and a test sample. The problem is that it wastes valuable resources if the total sample is not sufficiently large. A common practice is to perform cross validation (CV). A typical CV is to partition the sample into 10 parts of roughly equal size. We then use 9 parts to train the model and predict the remaining part. Eventually, all parts are predicted using samples that exclude the parts to be predicted. Such an approach is called a 10-fold cross validation. The result will depends on how the 10 parts are partitioned. Therefore, we often repartition the sample several times (say 20) and take the average predictability of the 20 replicates as the predictability. In generate, the method is called K-fold cross validation where K is the number of folds. K = 10 is the most preferred method but 5-fold is often adopted. The more the fold, the more the computational burden. When the sample size is small, n-fold cross validation may be conducted, where n is the sample size. This approach is also called leave-one-out-cross-validation (LOOCV), which represents the optimal approach because (1) it provides the highest predictability and a predictability closest to the true situation; (2) there is not sample partitioning error (only one possible partitioning). Figure 2 sketches a 5-fold cross validation,

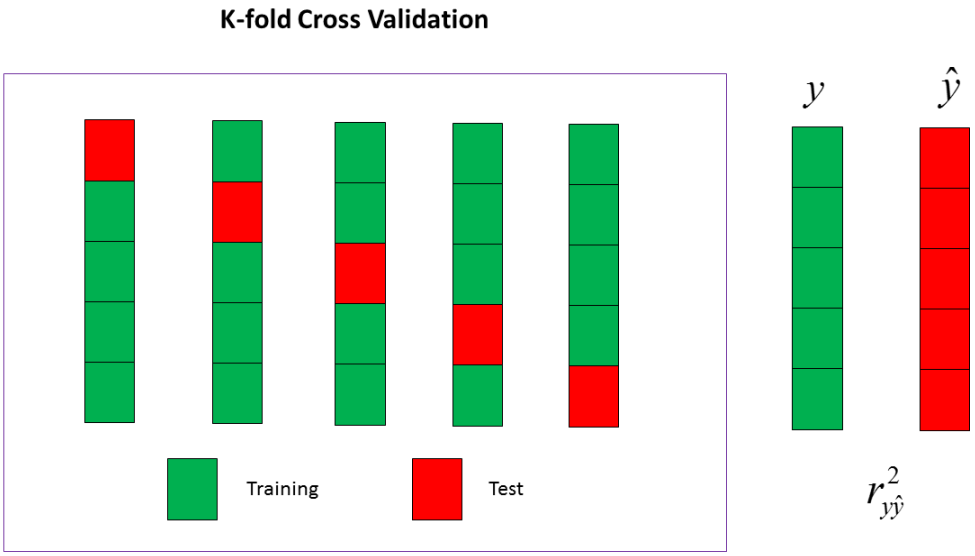


**Figure 2.** A five-fold cross validation where the parts labeled green are the training samples and labeled red are the test samples.

**Predictability in multiple linear regression analysis (fixed model)**

In linear regression analysis. The predictability can be evaluated without actually performing cross validation. There is an explicit formula to calculate the PRESS value using parameters estimated from the whole sample. Let

$$y = X\beta + e \tag{12}$$

be the linear model for a response variable predicted by a set of predictor *X*. The predicted residual error (not fitted error) for observation *j* is

$$\hat{y}_j = X_j \hat{\beta}_{[-j]} \tag{13}$$

where $\hat{\beta}_{[-j]}$ is a vector of estimated regression coefficients with the *j*th observation excluded. The predicted residual error for observation *j* is

$$e_j = y_j - \hat{y}_j = y_j - X_j \hat{\beta}_{[-j]} \tag{14}$$

Let

$$\hat{e}_j = y_j - X_j \hat{\beta} \tag{15}$$

be the estimated (fitted) residual error. There is an explicit relationship between the two errors (Cook 1977, 1979),

$$e_j = \frac{1}{1-h_{jj}} \hat{e}_j \tag{16}$$

where $h_{jj}$ is the leverage value of observation *j*, which is the corresponding diagonal element of the HAT matrix in defined below.,

$$H = X(X^T X)^{-1} X^T \tag{17}$$

The PRESS is calculated using

$$\text{PRESS} = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \frac{\hat{e}_j^2}{(1-h_{jj})^2} \tag{18}$$

The result is identical to the PRESS from the LOOCV. We extended Cook's method to leave $n_k$ out for weighted least squares regression analysis. The predicted *y* is a linear function of the observed *y* as shown below,

$$\hat{y} = X\hat{\beta} = X(X^T W X)^{-1} X^T W y = Hy \tag{19}$$

where

$$H = X(X^T W X)^{-1} X^T W \tag{20}$$

is the hat matrix. This *H* matrix is idempotent. In a K-fold cross validation analysis, let $n_k$ be the number of observations in the *k*th fold for $k = 1, \cdots, K$ and $\sum_{k=1}^{K} n_k = n$. Define $X_k$ as an $n_k \times p$ matrix of independent variables for individuals in the *k*th fold. The "leverage" value for the *k*th fold is defined as an $n_k \times n_k$ matrix,

$$H_{kk} = X_k (X^T W X)^{-1} X_k^T W_k \tag{21}$$

where $W_k$ is the $n_k \times n_k$ subset of matrix *W* corresponding to the *k*th fold. This matrix must appear in the end, not in the beginning, of the above equation. Let

$$\hat{e}_k = y_k - X_k \hat{\beta} \tag{22}$$

be the estimated residual errors where $\hat{\beta}$ is estimated from the whole sample. The predicted residual errors for the $n_k$ individuals in the *k*th fold is

$$e_k = (I - H_{kk})^{-1} \hat{e}_k \tag{23}$$

Therefore, the PRESS is defined as

$$\text{PRESS} = \sum_{k=1}^{K} e_k^T W_k e_k = \hat{e}_k^T (I - H_{kk})^{-1} W_k (I - H_{kk})^{-1} \hat{e}_k \tag{24}$$

which is the weighted sum of squares of the predicted residual errors. Derivation of equation (33) is given in **Appendix A** of the attached manuscript by Xu (2017).

**Predictability in mixed model analysis (Xu 2017)**

This part can be found in the attached manuscript and its derivation is given in **Appendix B** of the attached manuscript.

# Predicted Residual Error Sum of Squares of Mixed Models – An Application to Genomic Prediction

**Shizhong Xu**
Department of Botany and Plant Sciences, University of California, Riverside, CA 92521
E-mail: shizhong.xu@ucr.edu

## Abstract

Genomic prediction is a statistical method to predict phenotypes of polygenic traits using high throughput genomic data. Most diseases and behaviors in human and animals are polygenic traits. The majority of agronomic traits in crops are also polygenic. Accurate prediction of these traits can help medical professionals diagnose acute diseases and breeders to increase food products, and therefore, significantly contribute to human health and global food security. The best linear unbiased prediction (BLUP) is an important tool to analyze high throughput genomic data for prediction. However, to judge the efficacy of the BLUP model with a particular set of predictors for a given trait, one has to provide an unbiased mechanism to evaluate the predictability. Cross-validation (CV) is an essential tool to achieve this goal, where a sample is partitioned into *K* parts of roughly equal size, one part is predicted using parameters estimated from the remaining K – 1 parts and eventually every part is predicted using a sample excluding that part. Such a cross-validation is called the K-fold CV. Unfortunately, CV presents a substantial increase in computational burden. We developed an alternative method, the HAT method, to replace CV. The new method corrects the *estimated residual errors* from the whole sample analysis using the leverage values of a hat matrix of the random effects to achieve the *predicted residual errors*. Properties of the HAT method were investigated using seven agronomic and 1000 metabolomic traits of an inbred rice population. Results showed that the HAT method is a very good approximation of the CV method. The method was also applied to ten traits in 1495 hybrid rice with 1.6 million SNPs and to human height of 6161 subjects with roughly 0.5 million SNPs of the Framingham heart study data. Predictabilities of the HAT and CV methods were all similar. The HAT method allows us easily to evaluate the predictabilities of genomic prediction for large number of traits in very large populations.

KEY WORDS: Best linear unbiased prediction; Cross validation; Generalized cross validation; Genomic selection; Hybrid breeding; Mixed model.

## Author Summary

Cross-validation (CV) is an essential tool to evaluate the predictability of a prediction model. Unfortunately, CV presents a substantial increase in computational burden. We developed an alternative method, the HAT method, to replace CV. The new method corrects the *estimated residual errors* from the whole sample analysis using the leverage values of a hat matrix of the random effects to achieve the *predicted residual errors*. Properties of the HAT method were investigated using seven agronomic traits and 1000 metabolomic traits of an inbred rice population. Results showed that the HAT method is a very good approximation of the CV method.

## Introduction

Many diseases, anatomic structures, physiological characteristics and behaviors in human are polygenic traits. Most agronomic traits in agriculture, e.g., yield, are also polygenic. These complex traits require whole genome study to understand the genetic mechanisms and to genetically improve the quality and quantity of agricultural products (DE LOS CAMPOS *et al.* 2009; DE LOS CAMPOS *et al.* 2013a; DE LOS CAMPOS *et al.* 2013b). Genomic prediction (selection) is a statistical method of whole genome study (MEUWISSEN *et al.* 2001). It can lead to earlier detection of acute polygenic cancers (VAZQUEZ *et al.* 2012). Genomic prediction is also an effective tool to select superior cultivars in plant breeding (HEFFNER *et al.* 2009). Genomic hybrid prediction will provide an opportunity to evaluate all potential hybrids and allow breeders to select superior hybrids that will have little chance to be discovered based on traditional hybrid breeding schemes (XU *et al.* 2014). Genomic selection has been very successful in dairy cattle industry (GODDARD AND HAYES 2007) and will soon become a routine procedure of breeding for a vast number of agricultural species.

Among the commonly used methods for genomic prediction, the best linear unbiased predictor (BLUP) (HENDERSON 1975) is one of a few suitable methods for handling high throughput genomic data with millions of genetic variants (VANRADEN 2008). Reproducing kernel Hilbert spaces (RKHS) regression (GIANOLA *et al.* 2006) is another method with such an ability but RKHS has not been as well recognized as the BLUP method. Although variable selection approaches such as Bayes B (MEUWISSEN *et al.* 2001) and LASSO (TIBSHIRANI 1996) are optimal for traits with a few detectable loci of large effects plus many undetectable modifying loci under low and intermediate marker density, BLUP is the most robust method and one of the most commonly used genomic selection methods (DE LOS CAMPOS *et al.* 2013b). More importantly, the computational speed does not depend on marker density because it takes a marker inferred kinship matrix (covariance structure) as the input data, albeit computing kinship matrix taking additional time. To evaluate the predictability of the BLUP model, one has to resort to some other tools, such as validation or cross-validation, where individuals predicted do not contribute to estimated parameters that are used to predict these individuals. If individuals predicted are not excluded from the training sample, serious bias will occur in prediction.

The predictability of a model is often represented by the squared correlation coefficient between the observed and predicted phenotypic values (XU *et al.* 2014). This squared correlation is approximately equal to $R^2 = 1 - \text{PRESS}/\text{SS}$ where PRESS is the predicted residual error sum of squares and SS is the total sum of squares of the phenotypic values. Allen (ALLEN 1971; ALLEN 1974) proposed to use PRESS as a criterion to evaluate a regression model, in contrast to using the estimated residual sum of squares (ERESS) as the criterion. To calculate PRESS, Allen (ALLEN 1971; ALLEN 1974) used an approach that is now called the leave-one-out cross validation (LOOCV) or ordinary cross-validation (CRAVEN AND WAHBA 1979), in which an individual is predicted using parameters estimated from the sample that excludes this individual. When the sample size ($n$) is large, LOOCV presents a high computational cost because one will virtually have to analyze the data $n$ times. The K-fold cross validation (PICARD AND COOK 1984) is an extension of LOOCV in which the sample is partitioned into $K$ parts of roughly equal size. Individuals in a part are predicted simultaneously using all individuals in the remaining $K - 1$ parts. Eventually, all parts are predicted once and used to estimate parameters $K - 1$ times. When $K$ is small, there are many different ways of partitioning the sample, leading to variation in the calculated predictability. This variation can be very large for small sample sizes. Therefore, people often repeat the K-fold cross validation a few times and use their average values to reduce the error due to random partitioning. If possible, LOOCV (also called the n-fold cross validation, a special case of K-fold CV when $K = n$ and $n$ is the sample size) is recommended because it eliminates all problems associated with this random partitioning variation. However, such a cross validation is not realistic for large samples under the mixed model methodology. Although a simple split CV (50% training and 50% test) should suffice with very large samples, still 50% of the sample is wasted. The LOOCV method may slightly over predicts the model compared with the K-fold cross-validation when K is substantially smaller than $n$ (HASTIE *et al.* 2008).

Cook (COOK 1977; COOK 1979) developed an explicit method to calculate PRESS by correcting the deflated residual error of an observation using the leverage value of the observation without repeated analyses of the partitioned samples. This method applies to least square regression under the fixed model framework, where the predicted $y$ is a linear function of the observed $y$ as shown below,

$$\hat{y} = X\hat{\beta} = X(X^T X)^{-1} X^T y = Hy \tag{25}$$

where

$$H = X(X^T X)^{-1} X^T \tag{26}$$

is called the hat matrix. The predicted residual error for observation $j$ is $e_j = \hat{e}_j / (1 - h_{jj})$ where $\hat{e}_j = y_j - X_j\hat{\beta}$ is the so called estimated residual error and $h_{jj}$ is the leverage value of the $j$th observation (the $j$th diagonal element of the hat matrix). It is the contribution of the prediction for an individual from itself and may be called the *conflict of interest factor*. The predicted residual error is the estimated residual error after correction for the deflation. The sum of squares of the predicted residual errors over all individuals is the PRESS, which is a well-known statistic in multiple regression analyses. To find an explicit expression of PRESS for a mixed model, we need to identify a

random effect version of the hat matrix and use the leverage value of the *j*th observation to correct for the deflated residual error.

The HAT method is a fast algorithm for the ordinary cross-validation for a linear model (ALLEN 1971; ALLEN 1974) because the regression analysis is only done once on the whole sample and then the estimated residual errors are modified afterwards. Extension of the HAT method to mixed model has been made by Golab et al. (1979) in finding the optimal ridge factor in ridge regression (HOERL AND KENNARD 1970b; HOERL AND KENNARD 1970a). It is well known that ridge regression can be formulated as a mixed model problem with the variance ratio replaced by a given ridge factor. Golab et al. (1979) proposed a generalized cross-validation (GCV) method to find the optimal ridge factor so that the generalized residual error variance is minimized. These authors showed that the GCV calculated residual error sum of squares is a rotation-invariant version of Allen's PRESS. The residual error variance obtained from the GCV method is equivalent to calculating the residual error variance by dividing the estimated residual error sum of squares by an "effective" degree of freedom. Properties of the GCV method have been extensively studied by Li (1987). Jansen et al. (1997) applied GCV to wavelet thresholding. When performing genomic prediction, we prefer to see the actual predicted residual errors (errors in prediction of future individuals) obtained from the ordinary CV because the residual errors obtained via GCV may not be intuitive to most of us. The important gain from the GCV method of ridge regression analysis to genomic selection is the HAT matrix of the random model when the genomic variance is given.

There is a rich literature in smoothing spline analysis that also help us develop the fast HAT method for evaluation of mixed model predictability (WAHBA 1975; WAHBA AND WOLD 1975a; WAHBA AND WOLD 1975b; CRAVEN AND WAHBA 1979; WAHBA 1980; WAHBA 1990; WAHBA *et al.* 1995; WAHBA AND LUO 1997; WAHBA 1998; WANG 1998b; WANG 1998a; WAHBA *et al.* 2000; HASTIE *et al.* 2008). In smoothing spline curve fitting, a response variable is fit to a predictor with an arbitrary functional relationship. The common approach is to fit the curve using B-spline or other type of non-parametric approaches. Several spline bases (more than necessary) are constructed from the original predictor. These bases are considered as new predictors, which are then used to fit the response variable with linear relationship. The regression coefficients are then estimated using a penalized shrinkage method such as the ridge regression. The ridge parameter in smoothing splines is then called the smoothing parameter $(\lambda)$, which is often found so that the GCV residual error variance is minimized (CRAVEN AND WAHBA 1979; WAHBA 1980). Given the smoothing parameter, the predicted responses of all individuals are linear functions of all observed responses. Hastie et al. (2008) called these linear functions collectively the smoother matrix and denoted it by $S_\lambda$. This smoother matrix is the random effect version of the HAT matrix,

$$H^R = X(X^T X + \lambda Q)^{-1} X^T \qquad (27)$$

where *Q* is a known diagonal matrix. A HAT matrix under the random model was also given by de los Campos (DE LOS CAMPOS *et al.* 2013b) in the form of $\hat{y} = (G + \lambda I)^{-1} y = Hy$, although it was not derived for calculating PRESS. The HAT matrix of the fixed model introduced in equation (26) is then denoted by $H^F$. The

difference between the two HAT matrices is clear in form. Hastie et al. (2008) stated that both $H^R$ and $H^F$ are symmetric and positive semidefinite, $H^F$ is idempotent ($H^F H^F = H^F$) but $H^R$ is not, and $H^F$ has a rank of *m* (number of predictors) while $H^R$ has a rank of *n* (number of observations). So, the HAT matrix for a random model has been defined by the smoothing splines community. We may implement this HAT matrix in our BLUP prediction to evaluate the predictability of our models and avoid the lengthy CV analysis. The smoothing parameter (our variance ratio) should be given a reasonable value and the REML estimate from the whole sample is a natural choice. However, replacing the prechosen $\lambda$ by a data-driven estimate makes the HAT matrix a complicated function of the data. The question is What is the difference between the HAT method (when $\lambda$ is estimated from the whole sample) and the actual CV (when $\lambda$ is estimated anew within each fold)? This becomes the main objective of this study.

When revising this manuscript, a similar study was published in the same journal (G3) by Gianola and Schon (2016). They also recognized the approximation nature of the new method and stated that using the whole-sample-estimated $\lambda$ in place of the prechosen $\lambda$ will not affect the result too much, especially when the LOOCV is compared because the training sample only differs from the whole sample by one observation. However, this is only a speculation (most likely true) and they did not explicitly investigate the difference. Since the new method represents a significant technical improvement in genomic selection, the community must be aware of the difference before widely adopting the new method to evaluate a genomic selection program. In this study, we explicitly answer this question by analyzing several agronomic traits and 1000 metabolomic traits from two rice populations. Further comparison was also made in genomic prediction of human height from the Framingham heart study data (DAWBER *et al.* 1951; DAWBER *et al.* 1963).

## Models and Methods

### Fixed model

The HAT method for calculating PRESS under the fixed model is given by Cook (COOK 1977; COOK 1979) for the LOOCV scenario but not for the leave $n_k$ out cross-validation (the K-fold CV). We extended Cook's method to leave $n_k$ out for weighted least squares regression analysis. The predicted *y* is a linear function of the observed *y* as shown below,

$$\hat{y} = X\hat{\beta} = X(X^T WX)^{-1} X^T Wy = Hy \tag{28}$$

where

$$H = X(X^T WX)^{-1} X^T W \tag{29}$$

is the hat matrix. This *H* matrix is still idempotent. In a K-fold cross validation analysis, let $n_k$ be the number of observations in the *k*th fold for $k = 1, \cdots, K$ and $\sum_{k=1}^{K} n_k = n$. Define $X_k$ as an $n_k \times p$ matrix of independent variables for individuals in the *k*th fold. The "leverage" value for the *k*th fold is defined as an $n_k \times n_k$ matrix,

$$H_{kk} = X_k (X^T W X)^{-1} X_k^T W_k \tag{30}$$

where $W_k$ is the $n_k \times n_k$ subset of matrix *W* corresponding to the *k*th fold. This matrix must appear in the end, not in the beginning, of the above equation. Let

$$\hat{e}_k = y_k - X_k \hat{\beta} \tag{31}$$

be the estimated residual errors where $\hat{\beta}$ is estimated from the whole sample. The predicted residual errors for the $n_k$ individuals in the *k*th fold is

$$e_k = (I - H_{kk})^{-1} \hat{e}_k \tag{32}$$

Therefore, the PRESS is defined as

$$\text{PRESS} = \sum_{k=1}^{K} e_k^T W_k e_k = \hat{e}_k^T (I - H_{kk})^{-1} W_k (I - H_{kk})^{-1} \hat{e}_k \tag{33}$$

which is the weighted sum of squares of the predicted residual errors. Derivation of equation (33) is given in **Appendix A**.

**Mixed model**

The linear mixed model for genomic prediction is written as

$$y = X\beta + \xi + e \tag{34}$$

where $X\beta$ represents the fixed effects, $\xi$ is a vector of random (polygenic) effects with an assumed $N(0, A\sigma_\xi^2)$ distribution, and $e \sim N(0, I\sigma^2)$ is a vector of residual errors. The expectation and variance of *y* are $\mathrm{E}(y) = X\beta$ and $\text{var}(y) = V = A\sigma_\xi^2 + I\sigma^2$, respectively, where *A* is a marker inferred kinship matrix (explained in detail below), $\sigma_\xi^2$ is the polygenic variance and $\sigma^2$ is the residual error variance. The parameters are $\theta = \left\{\beta, \sigma_\xi^2, \sigma^2\right\}$ and the variances are estimated using the restricted maximum likelihood method (PATTERSON AND THOMPSON 1971) by maximizing the following likelihood function,

$$L(\theta) = -\frac{1}{2}\ln|V| - \frac{1}{2}\ln|X^{\mathrm{T}} V^{-1} X| - \frac{1}{2}(y - X\beta)^{\mathrm{T}} V^{-1} (y - X\beta) \tag{35}$$

The estimated genomic heritability (DE LOS CAMPOS *et al.* 2015) from the markers is $\hat{h}^2 = \hat{\sigma}_\xi^2 / (\hat{\sigma}_\xi^2 + \hat{\sigma}^2)$. The best linear unbiased estimates (BLUE) of the fixed effects are $\hat{\beta} = (X^{\mathrm{T}} V^{-1} X)^{-1} X^{\mathrm{T}} V^{-1} y$ and the best linear unbiased prediction (BLUP) of the polygenic effects are $\hat{\xi} = \hat{\sigma}_\xi^2 A V^{-1} (y - X\hat{\beta})$. The fitted phenotypic values are $\hat{y} = X\hat{\beta} + \hat{\xi}$, which is a conditional prediction (not a marginal prediction). Corresponding to the predicted polygenic effect $\hat{\xi} = \hat{y} - X\hat{\beta}$, we now define $\xi = y - X\hat{\beta}$ as the "observed" polygenic effect (it is indeed observed because $\hat{\beta}$ is used). The model goodness of fit for the random effects is defined as the squared correlation between $\xi$ and $\hat{\xi}$.

**Marker inferred kinship matrix**

The marker inferred kinship matrix *A* is calculated from all markers of the genome using the following equation,

$$A = \frac{1}{a}\sum_{k=1}^{m} Z_k Z_k^{\mathrm{T}} \tag{36}$$

where *m* is the total number of markers, $a = n^{-1}\mathrm{tr}\left(\sum_{k=1}^{m} Z_k Z_k^{\mathrm{T}}\right)$ is a normalization factor to make the diagonal elements of matrix *A* as close to unity as possible and $Z_k$ is an $n \times 1$ vector of genotype indicator variables for all individuals at marker *k*. For individual *j*, the numerical code for a genotype is

$$Z_{jk} = \begin{cases} -1 & \text{for } A_1A_1 \\ 0 & \text{for } A_1A_2 \\ +1 & \text{for } A_2A_2 \end{cases} \tag{37}$$

where $A_1A_1$, $A_1A_2$ and $A_2A_2$ are the three genotypes of the marker. People often standardize the $Z_k$ vectors before using them to calculate the kinship matrix (see VanRaden 2008).

**Cross validation**

For a K-fold CV, we randomly partitioned the sample into *K* parts of roughly equal size. We then used $K-1$ parts to predict the remaining part. Let $y = \begin{bmatrix} y_k^{\mathrm{T}} & y_{-k}^{\mathrm{T}} \end{bmatrix}^{\mathrm{T}}$ be the vector of phenotypic values that are partitioned into $y_k^{\mathrm{T}}$ and $y_{-k}^{\mathrm{T}}$, where $y_k^{\mathrm{T}}$ is a vector of phenotypic values of all observations in the *k*th part (test sample) and $y_{-k}^{\mathrm{T}}$ is a vector of phenotypic values for all individuals excluding observations in the *k*th part (training sample). Corresponding to this partitioning of the sample, we have

$$\mathrm{E}(y) = \begin{bmatrix} X_k\beta \\ X_{-k}\beta \end{bmatrix} \tag{38}$$

and

$$\mathrm{var}(y) = V = \begin{bmatrix} V_{kk} & V_{k(-k)} \\ V_{(-k)k} & V_{(-k)(-k)} \end{bmatrix} = \begin{bmatrix} A_{kk}\sigma_\xi^2 + I\sigma^2 & A_{k(-k)}\sigma_\xi^2 \\ A_{(-k)k}\sigma_\xi^2 & A_{(-k)(-k)}\sigma_\xi^2 + I\sigma^2 \end{bmatrix} \tag{39}$$

The predicted phenotypic values in the test sample are

$$\mathrm{E}(y_k \mid y_{-k}) = \hat{y}_k = X_k\hat{\beta}_{-k} + \sigma_\xi^2 A_{k(-k)}(A_{(-k)(-k)}\sigma_\xi^2 + I\sigma^2)^{-1}(y_{-k} - X_{-k}\hat{\beta}_{-k}) \tag{40}$$

Let $\xi_k = y_k - X_k\hat{\beta}_{-k}$ be the "observed" polygenic effect (phenotypes of the test sample adjusted by the fixed effects or centered phenotypes) and

$$\hat{\xi}_k^{CV} = \hat{\sigma}_\xi^2 A_{k(-k)}(A_{(-k)(-k)}\hat{\sigma}_\xi^2 + I\hat{\sigma}^2)^{-1}(y_{-k} - X_{-k}\hat{\beta}_{-k}) \tag{41}$$

be the predicted polygenic effect for the test sample. After all parts of the sample are predicted, we calculate the predicted residual error sum of squares using

$$\text{PRESS} = \sum_{k=1}^{K} (\xi_k - \hat{\xi}_k^{CV})^{\mathrm{T}} (\xi_k - \hat{\xi}_k^{CV}) \tag{42}$$

The predictability is defined as

$$R^2_{\text{CV}} = 1 - \text{PRESS/SS} \tag{43}$$

where

$$\text{SS} = \sum_{k=1}^{K} (\xi_k - \bar{\xi})^{\mathrm{T}} (\xi_k - \bar{\xi}) \tag{44}$$

is the total sum of squares of *y* adjusted by the fixed effects.

**The HAT method**

With the HAT method, we first defined the adjusted or centered phenotypic vector by the fixed effects $\xi = y - X\hat{\beta}$ as the "observed" $\xi$ and define $\hat{\xi} = \hat{y} - X\hat{\beta}$ as the predicted $\xi$, where $\hat{\beta}$ is estimated from the whole sample. We then used the whole sample to predict the polygenic effects

$$\hat{\xi} = \hat{\sigma}^2_{\xi} A V^{-1} (y - X\hat{\beta}) = \hat{\sigma}^2_{\xi} A V^{-1} \xi \tag{45}$$

Comparing the second form of the above equation ($\hat{\xi} = \hat{\sigma}^2_{\xi} A V^{-1} \xi$) with the fixed model HAT function, $\hat{y} = Hy$, we realized that $\hat{\xi} = H^{\mathrm{R}} \xi$, where $H^{\mathrm{R}} = \hat{\sigma}^2_{\xi} A V^{-1}$ is the HAT matrix of the random effects. Substituting $V^{-1}$ by $(A\sigma^2_{\xi} + I\sigma^2)^{-1}$ and after a few steps of algebraic derivation leads to

$$H^{\mathrm{R}} = \hat{\sigma}^2_{\xi} A (A\sigma^2_{\xi} + I\sigma^2)^{-1} = (I + \lambda A^{-1})^{-1} \tag{46}$$

where $\lambda = \sigma^2 / \sigma^2_{\xi}$ is the variance ratio. With eigen-decomposition for the *A* matrix, we have $A = UDU^{\mathrm{T}}$, $A^{-1} = UD^{-1}U^{\mathrm{T}}$ and $UU^T = U^T U = I$. Therefore,

$$H^{\mathrm{R}} = U(I + D^{-1}\lambda)^{-1} U^{\mathrm{T}} = U(U^T U + \lambda D^{-1})^{-1} U^{\mathrm{T}} \tag{47}$$

This expression (the second form) is exactly the one defined by Hastie et al. (2008) for the smoothing spline analysis given in equation (27), where their *X* is replaced by the eigenvector *U*, their *Q* is replaced by the inverse of eigenvalue matrix $D^{-1}$ (diagonal), and their smoothing parameter is our variance ratio. The HAT matrix is easy to compute because $(I + \lambda D^{-1})^{-1}$ is diagonal. When some eigenvalues are zero, $D^{-1}$ does not exist (very often), we reformulate it by $D(D + \lambda I)^{-1}$. Therefore,

$$(I + \lambda D^{-1})^{-1} = D(D + \lambda I)^{-1} = \text{diag}\left\{ \frac{\delta_1}{\delta_1 + \lambda}, \frac{\delta_2}{\delta_2 + \lambda}, \cdots, \frac{\delta_n}{\delta_n + \lambda} \right\} \tag{48}$$

where $\delta_j$ is the *j*th eigenvalue of matrix *A*.

Although the HAT method does not need to refit the model for each part predicted, it still needs to partition the sample into *K* parts if comparison with the traditional CV is of interest. Let $\hat{e}_k = \xi_k - \hat{\xi}_k$ be the estimated residual errors for all individuals in the *k*th part and $H^{\mathrm{R}}_{kk}$ be the diagonal block of matrix $H^{\mathrm{R}}$ corresponding to all individuals in the *k*th

part. The predicted residual errors for the *k*th part are $e_k = (I - H^{\mathrm{R}}_{kk})^{-1}\hat{e}_k$. Proof of this predicted residual error is provided in **Appendix B**. The PRESS under this random model becomes

$$\mathrm{PRESS} = \sum_{k=1}^{K} e_k^{\mathrm{T}} e_k = \sum_{k=1}^{K} \hat{e}_k^{\mathrm{T}} (I - H^{\mathrm{R}}_{kk})^{-2} \hat{e}_k \tag{49}$$

The predictability is measured by

$$R^2_{\mathrm{HAT}} = 1 - \mathrm{PRESS}/\mathrm{SS} \tag{50}$$

where $\mathrm{SS} = \sum_{k=1}^{K} (\xi_k - \bar{\xi})^T (\xi_k - \bar{\xi})$ is the total sum of squares for the centered *y* (adjusted by the fixed effects). The n-fold HAT approach is a special case where the *k*th part to be predicted contains only one individual, i.e., $H^{\mathrm{R}}_{kk} = h^{\mathrm{R}}_{jj}$ for $k = j$. Therefore, the leave-one-out version of the PRESS is

$$\mathrm{PRESS} = \sum_{j=1}^{n} e_j^2 = \sum_{j=1}^{n} \hat{e}_j^2 / (1 - h^{\mathrm{R}}_{jj})^2 \tag{51}$$

This predictability is roughly equal to the squared correlation between the fixed-effect-adjusted phenotypes and the predicted polygenic effects. The estimated residual error sum of squares is $\mathrm{ERESS} = \sum_{j=1}^{n} \hat{e}_j^2$ and the usual R-square reported in regression analysis is $R^2 = 1 - \mathrm{ERESS/SS}$, which is a measurement of model goodness of fit, not predictability.

**Generalized cross-validation (GCV)**

Generalized cross-validation (GOLAB *et al.* 1979) is an alternative method to correct the deflated residual error variance. The GCV calculated residual error sum of squares is called generalized residual error sum of squares (GRESS), which is defined by

$$\mathrm{GRESS} = \frac{(\xi - \hat{\xi})^T (\xi - \hat{\xi})}{\left[ n^{-1} \mathrm{tr}(I - H^R) \right]^2} \tag{52}$$

where $\hat{\xi}$ is the predicted polygenic effect from the whole sample. It is equivalent to dividing each estimated residual error by the average $(1 - h_{jj})$ across all observations. Therefore, an intuitive expression of the above equation is

$$\mathrm{GRESS} = \sum_{j=1}^{n} \hat{e}_j^2 / (1 - \bar{h})^2 \tag{53}$$

where $\bar{h} = \sum_{j=1}^{n} h_{jj} / n$ is the average leverage value across all observations and $\hat{e}_j = \xi_j - \hat{\xi}_j$. The predictability is defined as

$$R^2_{\mathrm{GCV}} = 1 - \mathrm{GRESS/SS} \tag{54}$$

Golab et al. (1979) stated that GRESS is a rotation-invariant PRESS. It is not intuitive to interpret GRESS and therefore we prefer to report PRESS and thus $R^2_{\mathrm{HAT}}$.

# Results

## Properties of the HAT method

Properties of the HAT method will be demonstrated using an experimental rice population consisting of 210 recombinant inbred lines (YU *et al.* 2011). These lines were derived from the cross of two rice varieties. A total of 270,820 SNPs were used to infer breakpoints of the genome for each line, resulting in a total of 1619 bins. A bin is a haplotype block within which there are no breakpoints across the entire population. In the original analysis of Yu et al. (2011), each bin was treated as a genetic marker. In this study, we used all the 1619 bins to infer a 210 × 210 kinship matrix. The matrix represents the genetic relationships of the lines and is used to model the covariance structure of the polygene. The population size is reasonably small and enabled us to compare the HAT method with CV in great detail.

Seven agronomic and 1000 metabolomic traits were included in the analysis. The agronomic traits are yield per plant (YD), tiller number per plant (TP), grain number per panicle (GN), 1000-grain weight (KGW), grain length (GL), grain width (GW) and heading day (HD). The first four traits (YD, TP, GN, and KGW) were field evaluated four times (two locations in two years), and GL and GW were replicated twice (two different years), and HD was replicated three times (three different years). The phenotypic value of each trait for each line is the average of the replicates. The 1000 metabolites were measured from seeds (317) and leaves (683) with two biological replications (GONG *et al.* 2013). The phenotypic values of the metabolites are the average expression levels of the two replicates after log2 transformation.

Predictability of the HAT method was compared with that of the CV method starting at 2-fold and ending at n-fold incremented by one, as shown in **Figure 1** for the seven agronomic traits. The two methods produced very similar values of R-squares, with a slight upward bias for the HAT method due to the use of $\lambda$ estimated from the whole sample. The biases are quite small for high predictability traits, e.g., KGW and GL. They appear to be large for low predictability traits such as YD and HD. However, this is partly due to the small scale of the *y*-axis (a visual effect). For example, the predictabilities of HAT and CV for trait KGW are 0.7564 and 0.7534, respectively, and the corresponding predictabilities for trait HD are 0.0774 and 0.0653. **Figure 1** also shows that when the numbers of folds are small, the predictabilities vary wildly and the variation progressively reaches zero at n-fold. The variation is caused by the ways that the folds are partitioned within the sample. Therefore, when a low number of folds are used in CV, it is necessary to repeat the CV a few times to reduce this variation. Although multiple CV will cause extra computational time, the HAT method can easily evaluate this variation.

Since computing the HAT method is sufficiently fast, we were able to perform random partitioning of the sample 100 times within a few minutes for all folds running from 2 to *n*. The last panel of **Figure 1** shows the mean and 95% confidence band for the replicated HAT predictability for trait KGW. The average predictability reaches a plateau at about 10 folds, but the 95% band is still very wide. This result did not support the claim that the

LOOCV seriously biased the predictability compared with K-fold CV (HASTIE *et al.* 2008).

**Figure 2** (A) shows the plot of predictability from n-fold CV against that from HAT for the seven agronomic traits of rice. The differences between the two methods are visually indistinguishable. We then compared the two methods for the 1000 metabolomic traits with n-fold cross validation. The CV method took a few days to complete the n-fold cross validation but the HAT method, again, took no more than a few minutes. The corresponding plots for the 1000 metabolomic traits are shown in **Figure 2** (B). Except three outliers, all points fall on the diagonal line. The three outliers show that the HAT prediction is over optimistic.

**Genomic hybrid prediction in rice**

We used a hybrid population of rice (HUANG *et al.* 2015) to demonstrate the application of the HAT method to genomic hybrid breeding. The population consists of 1495 hybrid rice with 10 agronomic traits measured in two locations in China (Hangzhou and Sanya). The 10 traits are grain yield (YD), panicle number (PN), grain number (GN), seed setting rate (SSR), 1000 grain weight (KGW), heading date (HD), plant height (PH), panicle length (PL), grain length (GL) and grain width (GW). The phenotypic value of each hybrid is the average of the two locations. We used 1.6 million SNPs to infer the kinship matrix and then performed predictions using both the HAT and CV methods. Although the n-fold sample partitioning can be easily accomplished with the HAT method, it would be too costly to do it with the CV method. Therefore, we compared the two methods under the 10-fold cross validation. We replicated the experiment 20 times per 10-fold CV to reduce the variation caused by random partitioning of the sample. The average of the 20 replicates presents the predictability for each method.

The two replications of hybrid rice experiments allowed us to estimate trait heritability of the hybrid population using the traditional ANOVA method (**Table 1**). We partitioned the phenotypic variance into variance due to hybrids (genetic) and variance due to residual error with systematic difference between the two locations excluded from the phenotype.

First, we compared the predictability of the CV method with the HAT method. **Figure 2** (C) shows the plot of the CV generated predictability against the HAT generated predictability. All the ten points (one point per trait) fall on the diagonal line, indicating very good agreement between the two methods. We then compared the trait heritability (H2) from the two replicated environments with the predictability drawn from 10-fold cross validation (CV), the predictability obtained from the HAT method (HAT) and the goodness of fit (FIT). The plots are illustrated in **Figure 2** (D). The $R^2$ of HAT and CV are the same (the red circles overlap with the blue circles). Both HAT and CV fall around the diagonal line with some upward biases compared to H2. The goodness of fit (FIT) are severely biased upwards and are not good representatives of H2 at all.

**Figure 3** shows a side-by-side comparison of H2 (trait heritability), $R^2$ of HAT, CV and FIT for all ten traits, where FIT is equivalent to genomic heritability (DE LOS CAMPOS *et al.* 2015). Different traits have very different H2, ranging from 0.08 (YD) to 0.92 (GL). The difference between HAT and CV is virtually zero across all traits and both are higher than H2 for the majority of the traits. For the three highly heritable traits (KGW, GL and GW), the H2 is higher than or equal to HAT and CV. Interestingly, HAT and CV are substantially higher than H2 for HD.

**Prediction of human height**

We analyzed human height of 6161 subjects from the Framingham heart study (DAWBER *et al.* 1951; DAWBER *et al.* 1963) with approximately 0.5 million SNPs using the mixed model methodology incorporating marker inferred kinship matrix. The model included effects of generation (two levels) and gender (male and female) as fixed effects. The estimated polygenic and residual variances are $\hat{\sigma}^2_{\xi} = 9.2375$ and $\hat{\sigma}^2 = 1.2617$, respectively, yielding a $\hat{\lambda} = 0.1365897$ and an estimated genomic heritability of $\hat{h}^2 = 0.8798$. This genomic heritability is close to the reported gender average heritability of human height (0.75-0.88) (SILVENTOINEN *et al.* 2003). The 10-fold CV and the HAT method gave predictabilities of $0.3063 \pm 0.0079$ and $0.3151 \pm 0.0037$, respectively. Note that the predictabilities are the averages of 20 replicated random partitions and thus there are small standard errors associated with the average values. The predictability obtained from the leave-one-out HAT method is 0.3278, slightly higher than the 10-fold partitioning approach.

**Generalized cross-validation and optimization of $\lambda$**

Before we perform the following analysis, it is worthwhile to refresh our mind that the HAT method will slightly over estimate the predictability because of the approximation nature. We first used the human height trait as an example to demonstrate the difference between the HAT method and the GCV method. The REML estimate of the variance ratio is $\hat{\lambda} = 0.1366$ and the corresponding predictability from the n-fold HAT method is $R^2_{\text{HAT}} = 0.3278$. This REML estimate generates a GCV predictability of $R^2_{\text{GCV}} = 0.3536$, different from that of the HAT method. We now treated $\lambda$ as a tuning parameter to maximize the predictability, as done by Mathew et al. (2015) in GCV for estimating breeding values. Using grid search around the REML estimated value ($\hat{\lambda} = 0.1366$), we found that the maximum achievable predictability for the HAT method is $R^2_{\text{MAX}} = 0.3310$ when $\lambda = 0.218$, leading to a gain of $0.3310 - 0.3278 = 0.0032$, which represents a $(0.3310\text{-}0.3278)/0.3278 \approx 1\%$ gain in predictability. Although this gain is negligible, it demonstrates that the REML estimated parameter does not give the maximum predictability. The good news is that $\hat{\lambda}$ is almost optimal, at least in this example. The corresponding maximum achievable predictability in GCV is $R^2_{\text{MAX}} = 0.3539$ when $\lambda = 0.158$, leading to a gain of $0.3539 - 0.3536 = 0.0003$. **Figure 4** shows the predictability profiles around $\hat{\lambda} = 0.1366$. By tuning the parameter, the gain in

predictability of the HAT method (panel A) is visible but the gain of the GCV method (panel B) is not recognizable.

To further compare the predictabilities of the HAT and GCV methods with their maximum achievable predictabilities, we used the "Brent" method of the "opim()" function in R to search for the optimal tuning parameter ( $\lambda$ ) for all 1000 metabolomic traits in the inbred rice population (210 lines). These optimal values of $\lambda$ may be called the maximum predictability estimates (MPE). **Figure 5** illustrates the comparisons of predictabilities across all 1000 traits, where more than a dozen traits show visible gains in predictability by tuning the parameter around the REML estimated value for the HAT method (**Figure 5**, panel A). Similar comparison is shown in **Figure 5** (panel B) for the GCV method where tuning the parameter achieves more than 20 visible gains in predictability. **Figure 5** (panel C) compares the predictabilities of GCV and HAT when the tuning parameter is fixed at the REML estimated value. The two methods provided very similar predictabilities for all 1000 traits except a half dozen traits with visible differences. All three comparisons shown in **Figure 5** have fitted R-squares at about 0.9995 and the regression coefficients are not significantly different from one ($p > 0.05$) except panel C where the regression coefficient is significantly greater than one ($p < 0.05$).

## Discussion

Very recently, Gianola and Schon (2016) published methods very similar to our HAT method to evaluate the predictability of a genomic selection model. They also recognized the approximation nature of the method when the smoothing parameter $\lambda$ is replaced by the estimated value from the whole sample. Their justification of the use of this whole sample estimated parameter, particularly in LOOCV, is that the estimated $\lambda$ from the whole sample will not be much different from the ones obtained from the training samples that differ from the whole sample by just one observation. They actually investigated the variation of $\lambda$ across all training samples and found that the variance is indeed small. Gianola and Schon (2016) investigated the properties of the new methods in many different situations using an inbred population of wheat ($n$ = 599) to see how the predictability changes when the training and test sample size ratio changes. These exhaustive investigations would take months or years to complete if the ordinary cross-validation were carried out. In addition to BLUP, these authors also extended the method to RKHS (GIANOLA *et al.* 2006) and the Bayesian alphabetic series (GIANOLA 2013) by modifying the importance sampling schemes.

One important issue that was not addressed in Gianola and Schon (2016) is how much difference in predictability calculated between the fast method and the classical CV method can be expected. This question is fundamental because the new method represents a significant technical improvement in genomic selection and will be adopted widely soon after the GS community recognizes it. In our study, we particularly focused on this question and investigated the difference using seven traits from an inbred population of rice, 1000 metabolomic traits from the same inbred population, 10 traits from a hybrid population of rice and one trait (human height) from a large human

population. We found that the HAT method always provides a slightly biased predictability over that of the CV method. However, the bias is never sufficiently severe to distort the conclusion on the predictability of a model. For example, in the human height prediction, the 10-fold CV produced a predictability of $0.3063$ and the corresponding number from the 10-fold HAT method was $0.3151$. The model goodness of fit, however, is 0.8789. The HAT method gave a number much closer to the CV predictability than the model goodness of fit.

In addition to comparing the differences between the HAT method and the ordinary CV, we also compared the new HAT method with the GCV method (GOLAB *et al.* 1979) and found that the two produced very similar results. Craven and Wahba (1979) compared GCV with CV and concluded that the smoothing parameter that maximizes the CV was amazingly close to the parameter that maximizes GCV. The GCV method has been available for almost four decades, but the genomic selection community, except Mathew et al. (2015), has never paid attention to it. Our study showed that both GCV and HAT can be applied to genomic selection. However, the HAT method directly addresses prediction of future individuals and therefore it is more intuitive to interpret the result.

Hastie et al. (2008) claimed that LOOCV provides a biased prediction compared with CV with lower number of folds. We observed that when the number of folds is 10 or above, the predictability stabilizes (**Figure 1**, the last panel). We did not observe a progressive increase of the predictability as the number of folds increases. Therefore, from our study, we recommend to perform LOOCV with the HAT method to avoid variation caused by random partitioning of the samples when the number of folds is small. When 10-fold or 5-fold CV is carried out, the analysis will only be conducted 10 or 5 times, which may not be a big deal and, therefore, the HAT method may loose its appeal. This statement may not be true considering the fact that the 10-fold CV must be run many times to reduce the variation caused by random partitioning of the samples. A multiple CV analysis for large samples is a significant burden to investigators. Therefore, the HAT method is a good alternative to CV to evaluate a genomic selection program.

We originally hoped to see a significant improvement in predictability by tuning the smoothing parameter around the REML estimated parameter. It is much disappointed that there were very few significant improvements from predictions of 1000 traits. The largest improvement occurred for the 422th metabolite with an improvement of $(0.6683-0.5615)/0.5615=0.1902048\approx 19\%$ (see the red point most deviating away from the diagonal line in **Figure 5**, panel A). The good news is that in most cases, the REML estimate is close to the MPE and, therefore, the parameter does not need to be tuned. On the other hand, since the computation is simple and fast, why not go ahead to tune the parameter and, if lucky, we may get an improved predictability, like the 422th metabolite in the inbred rice population.

In mixed model prediction, the random effects are often the targets for prediction. This is the case in genomic prediction because the genetic values are treated as random effects. However, if the investigators are interested in prediction using the fixed effects only under the mixed model, the estimated marginal residual error need to be adjusted by the

leverage values from the fixed model hat matrix $H^{\mathrm{F}} = X(X^{\mathrm{T}}V^{-1}X)^{-1}X^{\mathrm{T}}V^{-1}$ (SCHABENBERGER 2004). Let $\hat{e}_k = y_k - X_k\hat{\beta}$ be the estimated marginal residual errors for individuals in the *k*th fold, the predicted marginal residual errors are approximated by $e_k = (I - H^{\mathrm{F}}_{kk})^{-1}\hat{e}_k$, where $H^{\mathrm{F}}_{kk}$ is the diagonal block of $H^{\mathrm{F}}$ corresponding to observations in the *k*th fold. The MIXED procedure in SAS calls this method the non-iterative influence diagnostics while the iterative influence diagnostics is through actual cross validation (refit model and re-estimate covariance parameters). The non-iterative and iterative influence diagnostics can be interpreted as the HAT method and the CV method, respectively. PROC MIXED does not provide influence diagnostics for prediction of random effects. If there is an interest in both the fixed and random effects for prediction, the HAT matrix should include both the fixed model part and the random model part of the HAT matrix, $H^{\mathrm{M}} = H^{\mathrm{F}} + H^{\mathrm{R}}(I - H^{\mathrm{F}})$. The estimated conditional residual errors are $\hat{e}_k = y_k - X_k\hat{\beta} - \hat{\xi}_k$ and the predicted conditional residual errors are obtained by $e_k = (I - H^{\mathrm{M}}_{kk})^{-1}\hat{e}_k$, where $H^{\mathrm{M}}_{kk}$ is the diagonal block of $H^{\mathrm{M}}$ corresponding to observations in the *k*th fold.

When the mixed model includes multiple covariance structures, say *S* covariance structures, a similar $H^{\mathrm{R}}$ matrix is used except that the $\sigma^2_\xi A$ and $V$ matrices in $H^{\mathrm{R}}$ are replaced by $G = \sum_{s=1}^{S} A_s\sigma^2_s$ and $V = \sum_{s=1}^{S} A_s\sigma^2_s + I\sigma^2$, respectively, where $A_s$ is the *s*th covariance structure and $\sigma^2_s$ is the corresponding variance. An example of the multiple variance component model is the model with non-additive variances that include dominance and epistasis (XU 2013). Gianola and Schon (2016) also extended the new method to handle multiple kernels.

The HAT method applies to fixed models (exact result) and linear mixed models (approximate result). Is it possible to extend the HAT method to LASSO and PLS (partial lest squares)? An approximate extension may be possible by fixing the shrinkage parameter, like the extension to BLUP, but there is no exact extension. To carry out that approximate extension, we need to find the HAT function of the predicted *y* on the observed *y*, e.g., $\hat{y} = H^{\mathrm{LASSO}}y$ and $\hat{y} = H^{\mathrm{PLS}}y$. In general, the HAT matrix is $H = \partial\hat{y}/\partial y$ (SCHABENBERGER 2004), a Jacobian matrix holding each derivative of a predicted quantity with respect to an observed response.

## Data Availability

All data analyzed in this study are previously published. Sources of these data are provided by the references cited in the text. The Framingham Heart Study Data were downloaded from NCBI dbGaP with an IRB number HS -11-159.

## Supplementary Materials

Data of the inbred rice population and the R codes analyzing the data are available in **Supplementary Materials** associated with this article. **File S1** (mixedFunction.R) is the R code defining a function called mixed(). This function is used to estimate parameters of the mixed model using either the ML or the REML method along with the eigen-decomposition algorithm. The parameters include fixed effects (beta), polygenic variance (va) and residual variance (ve). In all the R codes, the variance ratio is defined as $\lambda_{\text{CODE}} = \sigma_\xi^2 / \sigma^2$, which is the inverse of the one described in the text $\lambda_{\text{TEXT}} = \sigma^2 / \sigma_\xi^2$. **File S2** (mixedBlupFunction.R) is the R code defining a function called mixedBlup(). This function is used to estimate parameters of the mixed model and perform BLUP prediction. If heritability is provided, the mixedBlup() function will only perform prediction. **File S3** (mixedCV.R) is an R program to perform mixed model analysis and cross validation (CV) for the inbred rice population. It takes "RIL-phe.csv" and "RIL-kk.csv" as the input data and generates three data.frames as the outputs. You need to source the two R functions described in File S1 and S2. The first data.frame (fit) is the result of mixed model analysis, including estimated fixed effects (beta), polygenic variance (va) and residual variance (ve). The second data.frame (PRED) stores the predicted phenotypic values and the predicted random effects (polygenic effects) from the CV analysis along with the original observed phenotypes and other information. The third data.frame (PRESS) stores the PRESS value and the R2 (predictability) resulted from the CV analysis. You should use the sample data to test the program first before analyzing your own data. Be sure to define the paths of the input and output data correctly! **File S4** (mixedHAT.R) is an R program to perform mixed model analysis and HAT prediction for the inbred rice population. It takes "RIL-phe.csv" and "RIL-kk.csv" as the input data and generates three data.frames as the outputs. You need to source the R function described in File S1. The first data.frame (fit) is the result of the mixed model analysis, including estimated fixed effects (beta), polygenic variance (va) and residual variance (ve). The second data.frame (PRED) stores the predicted phenotypic values and the predicted random effects (polygenic effects) from the HAT method along with the original observed phenotypes and other information. The third data.frame (PRESS) stores the PRESS value and the R2 (predictability) resulted from the HAT prediction. Again, you should use the sample data to test the program before analyzing your own. **File S5** (RIL-phe.csv) is a data set with 210 rows (excluding the heater) and nine columns. The first column is the ID of the lines. The second column is the fold ID used for cross validation analysis. Columns 3 – 9 are the phenotypic values of seven traits of the inbred rice. **File S6** (RIL-kk.csv) is the kinship matrix calculated from the whole genome marker data. This file has 210 rows (excluding the header) and 212 columns. The first and second columns are required by the MIXED procedure of SAS but are not needed by the R programs. In the R codes that require the kinship matrix, the first two columns should be removed prior to the analysis. The kinship matrix should be an 210×210 symmetric matrix.

## Acknowledgements

The author appreciated two anonymous reviewers and the associate editor for their constructive comments and suggestions on the first version of the manuscript. The author also was grateful to Dr. Yanru Cui (postdoc) and Yang Xu (student) for their help in calculating the marker inferred kinship matrices for the hybrid rice and human populations. The project was supported by the National Science Foundation Collaborative Research Grant DBI-1458515 to SX.

# Literature Cited


Allen, D. M., 1971 Mean square error of prediction as a criterion for selecting variables. Technometrics 13**:** 469-475.

Allen, D. M., 1974 The relationship between variable selection and data agumentation and a method for prediction Technometrics 16**:** 125-127.

Cook, D., 1977 Detection of influential observation in linear regression. Technometrics 19**:** 15-18.

Cook, D., 1979 Influential observation in linear regression. Journal of American Statistical Association 74**:** 169-174.

Craven, P., and G. Wahba, 1979 Smoothing noisy data with spline functions: Estimating the correct degree of smoothing by the method fgeneralized cross-validation. Numerische Mathematika 31**:** 377-403.

Dawber, T. R., W. B. Kannel and L. P. Lyell, 1963 An approach to longitudinal studies in a community - Framingham Study. Annals of the New York Academy of Sciences 107**:** 539-&.

Dawber, T. R., G. F. Meadors and F. E. Moore, 1951 Epidemiological approaches to heart disease - the Framingham study. American Journal of Public Health and the Nations Health 41**:** 279-286.

de los Campos, G., J. M. Hickey, R. Pong-Wong, H. D. Daetwyler and M. P. L. Calus, 2013a Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics 193**:** 327–345.

de los Campos, G., H. Naya, D. Gianola, J. Crossa, A. Legarra *et al.*, 2009 Predicting quantitative traits with regression models for dense molecular markers and pedigree. Genetics 182**:** 375-385.

de los Campos, G., D. Sorensen and D. Gianola, 2015 Genomic heritability: What is it? PLos Genetics 11**:** e1005048.

de los Campos, G., A. I. Vazquez, R. Fernando, Y. C. Klimentidis and D. Sorensen, 2013b Prediction of complex human traits using the genomic best linear unbiased predicto. PLos Genetics 9**:** e1003608.

Gianola, D., 2013 Priors in whole-genome regression: the Bayesian alphabet returns. Genetics 194**:** 573-596.

Gianola, D., R. L. Fernando and A. Stella, 2006 Genomic-assisted prediction of genetic value with semiparametric procedures. Genetics 173**:** 1761-1776.

Gianola, D., and C.-C. Schon, 2016 Cross-validation without doing cross-validation in genome-enabled prediction. Genetics, Genomics and Genetics (G3) 6**:** 3107-3128.

Goddard, M. E., and B. J. Hayes, 2007 Genomic selection. Journal of Animal Breeding and Genetics 124**:** 323-330.

Golab, G. H., M. Heath and G. Wahba, 1979 Generalized cross-validation as a method for choosig a good ridge parameter. Technometrics 21**:** 215-223.

Gong, L., W. Chen, Y. Gao, X. Liu, H. Zhang *et al.*, 2013 Genetic analysis of the metabolome exemplified using a rice population. Proc. Natl Acad. Sci. USA 110**:** 20320–20325.

Hastie, T., R. Tibshirani and J. Friedman, 2008 *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Springer New York.

Heffner, E. L., M. E. Sorrells and J.-L. Jannink, 2009 Genomic selection for crop improvement. Crop Science 49**:** 1-12.

Henderson, C. R., 1975 Best linear unbiased estimation and prediction under a selection model. Biometrics 31**:** 423-447.

Hoerl, A. E., and R. W. Kennard, 1970a Ridge regression: applications to nonorthogonal problems. Technometrics 12**:** 69-82.

Hoerl, A. E., and R. W. Kennard, 1970b Ridge regression: Biased estimation for nonorthogonal problems. Technometrics 12**:** 55-67.

Huang, X., S. Yang, J. Gong, Y. Zhao, Q. Feng *et al.*, 2015 Genomic analysis of hybrid rice varieties reveals numerous superior alleles that contribute to heterosis. Nature Communication 6.

Jansen, M., M. Malfait and A. Bultheel, 1997 Generalized cross validation for wavelet thresholding. Signal Processing 56**:** 33-44.

Li, K.-C., 1987 Asymptotic optimality for $C_p$, $C_L$, Cross-validation and generalized cross-validation: discrete index set. The Annals of Statistics 15**:** 958-975.

Mathew, B., J. Le´on and M. J. Sillanpa¨a¨, 2015 Integrated nested Laplace approximation inference and cross-validation to tune variance components in estimation of breeding value. Molecular Breeding 35.

Meuwissen, T. H. E., B. J. Hayes and M. E. Goddard, 2001 Prediction of total genetic value using genome-wide dense marker maps. Genetics 157**:** 1819-1829.

Patterson, H. D., and R. Thompson, 1971 Recovery of inter-block information when block sizes are unequal. Biometrika 58**:** 545-554.

Picard, R. R., and D. Cook, 1984 Cross-validation of regression models Journal of the American Statistical Association 79**:** 575-583.

Schabenberger, O., 2004 Mixed model influence diagnostics, pp. in *SAS Institute Inc. 2004 Proceedings of the Twenty-Ninth Annual SAS® Users Group International Conference*, edited by SAS Institute Inc. SAS Institute Inc., Montréal, Canada.

Silventoinen, K., S. Sammalisto, M. Perola, D. I. Boomsma, B. K. Cornes *et al.*, 2003 Heritability of adult body height: A comparative study of twin cohorts in eight countries. Twin Research 6**:** 399-408.

Tibshirani, R., 1996 Regression shrinkage and selection via the Lasso. Journal of the Royal Statistical Society, Series B 58**:** 267-288.

VanRaden, P. M., 2008 Efficient methods to compute genomic predictions. Journal of Dairy Science 91**:** 4414-4423.

Vazquez, A. I., G. de los Campos, Y. C. Klimentidis, G. J. M. Rosa, D. Gianola *et al.*, 2012 A comprehensive genetic approach for improving prediction of skin cancer risk in humans. Genetics 192**:** 1493–1502.

Wahba, G., 1975 Smoothing noisy data with spline functions Numer. Math. 24**:** 383-393.

Wahba, G., 1980 Spline bases, regularization, and generalized crossvalidation for solving approximation problems with large quantities of noisy data, pp. 905-912 in *International Conference on Approximation Theory in Honour of George Lorenz*. Academic Press, Austin, Texas.

Wahba, G., 1990 *Spline Models for Observational Data*. SIAM, Philadelphia.

Wahba, G., 1998 Support vector machines, reproducing kernel Hilbert spaces and the randomized GACV, pp. University of Wisconsin, Madison WI.

Wahba, G., Y. Lin and H. Zhang, 2000 GACV for support vector machines, pp. 297–311 in *Advances in Large Margin Classifiers*, edited by A. Smola, P. Bartlett, B. Sch¨olkopf and D. Schuurmans. MIT Press, Cambridge, MA.

Wahba, G., and Z. Luo, 1997 Smoothing spline ANOVA fits for very large, nearly regular data sets, with application to historical global climate data. Ann. Numer. Math. 4**:** 579-598.

Wahba, G., Y. Wang, C. Gu, R. Klein and B. Klein, 1995 Smoothing spline ANOVA for exponential families, with application to the Wisconsin epidemiological study of diabetic retinopathy. The Annuals of Statistics 23**:** 1865-1895.

Wahba, G., and S. Wold, 1975a A completely automatic french curve: fitting spline functions by cross validation. Communications in Statistics 4**:** 1-17.

Wahba, G., and S. Wold, 1975b Periodic splines for spectral density estimation: The use of cross-validation for determining the degree of smoothing. Communications in Statistics 4**:** 125-141.

Wang, Y., 1998a Mixed effects smoothing spline analysis of variance. Journal of the Royal Statistical Society. Series B (Statistical Methodology) 60**:** 159-174.

Wang, Y., 1998b Smoothing spline models with correlated random errors. Journal of the American Statistical Association 93**:** 341-348.

Xu, S., 2013 Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195**:** 1209-1222.

Xu, S., D. Zhu and Q. Zhang, 2014 Predicting hybrid performance in rice using genomic best linear unbiased prediction. Proceedings of the National Academy of Sciences 111**:** 12456-12461.

Yu, H., W. Xie, J. Wang, Y. Xing, C. Xu *et al.*, 2011 Gains in QTL detection using an ultra-high density SNP map based on population sequencing relative to traditional RFLP/SSR markers. PLoS One 6**:** e17595. doi:17510.11371/journal.pone.0017595.

## Appendix A: Derivation of HAT prediction under the fixed model methodology

### 1. Predicted residual error sum of squares (PRESS)

In a K-fold cross validation analysis, let $n_k$ be the number of observations in the *k*th fold for $k = 1, \cdots, K$ and $\sum_{k=1}^{K} n_k = n$. Define $X_k$ as an $n_k \times p$ matrix of independent variables for the individuals in the *k*th fold. The "leverage" values for the *k*th fold is defined as an $n_k \times n_k$ matrix,

$$H_{kk} = X_k (X^T W X)^{-1} X_k^T W_k \tag{A1}$$

where $W_k$ is the $n_k \times n_k$ subset of matrix *W* corresponding to the *k*th fold. This matrix must appear in the end, not in the beginning, of the above equation. Let

$$\hat{e}_k = y_k - X_k \hat{\beta} \tag{A2}$$

be the estimated residual errors where $\hat{\beta}$ is estimated from the whole sample. The predicted residual errors for the $n_k$ individuals in the *k*th fold is

$$e_k = (I - H_{kk})^{-1} \hat{e}_k \tag{A3}$$

Therefore, the PRESS is defined as

$$\text{PRESS} = \sum_{k=1}^{K} e_k^T W_k e_k = \hat{e}_k^T (I - H_{kk})^{-1} W_k (I - H_{kk})^{-1} \hat{e}_k \tag{A4}$$

which is the weighted sum of squares of the predicted residual errors.

### 2. Derivation of predicted residual error sum of squares

The linear model for *p* independent variables (including the intercept) and *n* observations is

$$y = X\beta + e \tag{A5}$$

The weighted least squares estimates of all regression coefficients are obtained using

$$\hat{\beta} = (X^T W X)^{-1} X^T W y \tag{A6}$$

The estimated residual error variance is

$$\hat{\sigma}^2 = \frac{1}{n-p} (y - X\hat{\beta})^{\mathrm{T}} W (y - X\hat{\beta}) \tag{A7}$$

The variance-covariance matrix of the estimated regression coefficients are calculated using

$$\text{var}(\hat{\beta}) = (X^T W X)^{-1} \hat{\sigma}^2 \tag{A8}$$

which is an $p \times p$ matrix with diagonal elements being the variances and off-diagonal elements being the covariance.

The fitted values for all individuals in the population are

$$\hat{y} = X\hat{\beta} = X(X^{\mathrm{T}} W X)^{-1} X^{\mathrm{T}} W y \tag{A9}$$

Let us define a hat matrix by

$$H = X(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}W \tag{A10}$$

Therefore, the fitted values are a hat function of the observed values,

$$\hat{y} = Hy \tag{A11}$$

Let us partition the sample into *K* parts (folds) and denote the number of individuals in the *k*th fold by $n_k$. Define $y_k$ as an $n_k \times 1$ vector, which is a subset of *y* that contains all observations in the *k*th fold. Define $X_k$ as the $n_k$ rows of matrix $X$ corresponding to the observations in the *k*th fold. The predicted residual errors are

$$e_k = y_k - X_k\hat{\beta}_{(-k)} \tag{A12}$$

where

$$\hat{\beta}_{(-k)} = (X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)})^{-1}X_{(-k)}^{\mathrm{T}}W_{(-k)}y_{(-k)} \tag{A13}$$

are the estimated regression coefficients from the data with the $n_k$ observations in the *k*th fold being excluded. Let us make the following matrix decomposition,

$$X^{\mathrm{T}}WX = X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)} + X_k^{\mathrm{T}}W_kX_k \tag{A14}$$

Therefore,

$$X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)} = X^{\mathrm{T}}WX - X_k^{\mathrm{T}}W_kX_k \tag{A15}$$

Similarly, we can rewrite

$$X_{(-k)}^{\mathrm{T}}W_{(-k)}y_{(-k)} = X^{\mathrm{T}}Wy - X_k^{\mathrm{T}}W_ky_k \tag{A16}$$

Using Woodbury matrix identity (Woodbury 1950), we have

$$\begin{aligned}
&(X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)})^{-1} \\
&\quad = (X^{\mathrm{T}}WX - X_k^{\mathrm{T}}W_kX_k)^{-1} \\
&\quad = (X^{\mathrm{T}}WX)^{-1} - (X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}\left[X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}} - W_k^{-1}\right]^{-1}X_k(X^{\mathrm{T}}WX)^{-1} \\
&\quad = (X^{\mathrm{T}}WX)^{-1} + (X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k\left[I - X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k\right]^{-1}X_k(X^{\mathrm{T}}WX)^{-1} \\
&\quad = (X^{\mathrm{T}}WX)^{-1} + (X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k(I - H_{kk})^{-1}X_k(X^{\mathrm{T}}WX)^{-1}
\end{aligned} \tag{A17}$$

where

$$H_{kk} = X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k \tag{A18}$$

is an $n_k \times n_k$ matrix of leverage values for the *k*th fold. This matrix is the $n_k \times n_k$ diagonal block of the hat matrix *H*. Further derivation leads to

$$\begin{aligned}
&X_k(X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)})^{-1}X^{\mathrm{T}}Wy \\
&\quad = X_k(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy + X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k(I - H_{kk})^{-1}X_k(X^{\mathrm{T}}WX)^{-1}X^{\mathrm{T}}Wy \\
&\quad = X_k\hat{\beta} + H_{kk}(I - H_{kk})^{-1}X_k\hat{\beta}
\end{aligned} \tag{A19}$$

and

$$
\begin{aligned}
&X_k(X_{(-k)}^{\mathrm{T}}W_{(-k)}X_{(-k)})^{-1}X_k^{\mathrm{T}}W_k y_k \\
&= X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k y_k + X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k(I-H_{kk})^{-1}X_k(X^{\mathrm{T}}WX)^{-1}X_k^{\mathrm{T}}W_k y_k \\
&= H_{kk}y_k + H_{kk}(I-H_{kk})^{-1}H_{kk}y_k
\end{aligned}
\tag{A20}
$$

Therefore, the predicted residual errors are

$$
\begin{aligned}
e_k &= y_k - X_k\hat{\beta}_{(-k)} \\
&= y_k - X_k\hat{\beta} - H_{kk}(I-H_{kk})^{-1}X_k\hat{\beta} + H_{kk}y_k + H_{kkk}(I-H_{kk})^{-1}H_{kk}y_k \\
&= y_k + H_{kk}y_k + H_{kk}(I-H_{kk})^{-1}H_{kk}y_k - X_k\hat{\beta} - H_{kk}(I-H_{kk})^{-1}X_k\hat{\beta} \\
&= \left[I + H_{kk} + H_{kk}(I-H_{kk})^{-1}H_{kk}\right]y_k - \left[I + H_{kk}(I-H_{kk})^{-1}\right]X_k\hat{\beta}
\end{aligned}
\tag{A21}
$$

Note that

$$
(I-H_{kk})^{-1} = -H_{kk}^{-1} - H_{kk}^{-1}(I-H_{kk}^{-1})^{-1}H_{kk}^{-1} \tag{A22}
$$

Therefore,

$$
\begin{aligned}
H_{kk}(I-H_{kk})^{-1}H_{kk} &= H_{kk}(-H_{kk}^{-1} - H_{kk}^{-1}(I-H_{kk}^{-1})^{-1}H_{kk}^{-1})H_{kk} \\
&= -H_{kk} - (I-H_{kk}^{-1})^{-1}
\end{aligned}
\tag{A23}
$$

The coefficient of $y_k$ in equation (A21) is

$$
I + H_{kk} + H_{kk}(I-H_{kk})^{-1}H_{kk} = I + H_{kk} - H_{kk} - (I-H_{kk}^{-1})^{-1} = I + H_{kk}(I-H_{kk})^{-1} \tag{A24}
$$

which is identical to the coefficient of $X_k\hat{\beta}$ in equation (A21). Therefore,

$$
\begin{aligned}
e_k &= y_k - X_k\hat{\beta}_{(-k)} \\
&= \left[I + H_{kk} + H_{kk}(I-H_{kk})^{-1}H_{kk}\right]y_k - \left[I + H_{kk}(I-H_{kk})^{-1}\right]X_k\hat{\beta} \\
&= \left[I + H_{kk}(I-H_{kk})^{-1}\right]y_k - \left[I + H_{kk}(I-H_{kk})^{-1}\right]X_k\hat{\beta} \\
&= \left[I + H_{kk}(I-H_{kk})^{-1}\right](y_k - X_k\hat{\beta}) \\
&= \left[I + H_{kk}(I-H_{kk})^{-1}\right]\hat{e}_k
\end{aligned}
\tag{A25}
$$

We can see that the predicted residual errors are a linear function of the estimated residual errors. The next step is to simplify the linear function,

$$
\begin{aligned}
I + H_{kk}(I-H_{kk})^{-1} &= (I-H_{kk})^{-1}(I-H_{kk}) + H_{kk}(I-H_{kk})^{-1} \\
&= (I-H_{kk})^{-1}(I-H_{kk}+H_{kk}) \\
&= (I-H_{kk})^{-1}
\end{aligned}
\tag{A26}
$$

Therefore, the predicted residual errors have been expressed as a simple linear function of the estimated residual errors,

$$
e_k = (I-H_{kk})^{-1}\hat{e}_k \tag{A27}
$$

The predicted residual sum of squares (PRESS) is

$$
\mathrm{PRESS} = \sum_{k=1}^{K} e_k^T W_k e_k = \sum_{k=1}^{K} \hat{e}_k^T (I-H_{kk})^{-1}W_k(I-H_{kk})^{-1}\hat{e}_k \tag{A28}
$$

Let us define

$$
\Theta_{kk} = (I-H_{kk})^{-1}W_k(I-H_{kk})^{-1} \tag{A29}
$$

The PRESS is written as

$$\text{PRESS} = \sum_{k=1}^{K} \hat{e}_k^T \Theta_{kk} \hat{e}_k \tag{A30}$$

The PRESS is often translated into R-square to represent the predictability of a model,

$$\text{R}^2 = 1 - \frac{\text{PRESS}}{\text{SST}} \tag{A31}$$

where SST is the total sum of squares of the response variable.

**Reference**

Woodbury, M. A. 1950. Inverting modified matrices. *Memorandum Report, Statistical Research Group*, Princeton University, Princeton, NJ 42: 4-4.

## Appendix B: Proof of the HAT Method for PRESS in Mixed Models

### 1. Estimated random effects

Let us define

$$r = y - X\beta \tag{B1}$$

as the phenotypic values of the trait adjusted by the fixed effects, assuming that $\beta$ is known. The estimated random effects are more appropriately called the fitted random effects. Let us define the estimated vector of random effects by

$$\tilde{r} = K(K+\lambda I)^{-1} r = Hr \tag{B2}$$

where $H = K(K+\lambda I)^{-1}$ is the HAT matrix, $\lambda = \sigma^2 / \sigma_\xi^2$ is the variance ratio and $K$ is the kinship matrix. In the main text, we used *A* in place of *K*. Here we used *K* again to be consistent with the genomic selection literature. Let us define $\hat{e}_j = r_j - \tilde{r}_j$ as the estimated residual error for the *j*th observation or *j*th block of observations. The predicted residual error for the *j*th block of individuals is

$$e_j = (I - H_{jj})^{-1} \hat{e}_j \tag{B3}$$

The purpose of this appendix is to prove equation (B3) that the predicted residual error can be obtained from the estimated residual error via the leverage value (diagonal element or diagonal block) of the HAT matrix.

Let us partition the *K* matrix into

$$K = \begin{bmatrix} K_{jj} & K_{j(-j)} \\ K_{(-j)j} & K_{(-j)(-j)} \end{bmatrix} \tag{B4}$$

where $K_{jj}$ is the *j*th diagonal element of the *K* matrix, $K_{j(-j)}$ is the *j*th row of matrix *K* that excludes the *j*th column, and $K_{(-j)(-j)}$ is the *K* matrix excluding the *j*th row and the *j*th column. Corresponding to this partitioning, matrix $K+\lambda I$ can also be partitioned into

$$K + \lambda I = \begin{bmatrix} K_{jj} + \lambda I & K_{j(-j)} \\ K_{(-j)j} & K_{(-j)(-j)} + \lambda I \end{bmatrix} \tag{B5}$$

The inverse of the above partitioned matrix is denoted by

$$(K+\lambda I)^{-1} = \begin{bmatrix} K_{jj} + \lambda I & K_{j(-j)} \\ K_{(-j)j} & K_{(-j)(-j)} + \lambda I \end{bmatrix}^{-1} = \begin{bmatrix} C_{jj} & C_{j(-j)} \\ C_{(-j)j} & C_{(-j)(-j)} \end{bmatrix} \tag{B6}$$

where

$$\begin{aligned} C_{jj} &= \left[ (K_{jj} + \lambda I) - K_{j(-j)} (K_{(-j)(-j)} + \lambda I)^{-1} K_{(-j)j} \right]^{-1} \\ C_{j(-j)} &= -C_{jj} K_{j(-j)} (K_{(-j)(-j)} + \lambda I)^{-1} \\ C_{(-j)j} &= -(K_{(-j)(-j)} + \lambda I)^{-1} K_{(-j)j} C_{jj} \\ C_{(-j)(-j)} &= (K_{(-j)(-j)} + \lambda I)^{-1} + (K_{(-j)(-j)} + \lambda I)^{-1} K_{(-j)j} C_{jj} K_{j(-j)} (K_{(-j)(-j)} + \lambda I)^{-1} \end{aligned} \tag{B7}$$

The estimated (fitted) value of the *j*th individual is

$$\tilde{r}_j = \begin{bmatrix} K_{jj} & K_{j(-j)} \end{bmatrix} \begin{bmatrix} C_{jj} & C_{j(-j)} \\ C_{(-j)j} & C_{(-j)(-j)} \end{bmatrix} \begin{bmatrix} r_j \\ r_{-j} \end{bmatrix} \tag{B8}$$

which is eventually expressed as

$$\tilde{r}_j = K_{jj}C_{jj}r_j + K_{j(-j)}C_{(-j)j}r_j + K_{jj}C_{j(-j)}r_{-j} + K_{j(-j)}C_{(-j)(-j)}r_{-j} \tag{B9}$$

**2. Predicted random effects**

The predicted value for the *j*th individual is obtained by excluding the contribution from the same individual, as expressed below,

$$\hat{r}_j = K_{j(-j)}\left[K_{(-j)(-j)} + \lambda I\right]^{-1} r_{-j} \tag{B10}$$

Let us examine the four terms in the fitted value given in equation **Error! Reference source not found.**,

$$\begin{aligned} K_{jj}C_{jj}r_j &= K_{jj}C_{jj}r_j \\ K_{j(-j)}C_{(-j)j}r_j &= -K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}r_j \\ K_{jj}C_{j(-j)}r_{-j} &= -K_{jj}C_{jj}\hat{r}_j \\ K_{j(-j)}C_{(-j)(-j)}r_{-j} &= \hat{r}_j + K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}\hat{r}_j \end{aligned} \tag{B11}$$

Substituting these four terms into equation **Error! Reference source not found.**, we get

$$\begin{aligned} \tilde{r}_j = {} & K_{jj}C_{jj}r_j - K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}r_j \\ & -K_{jj}C_{jj}\hat{r}_j + \hat{r}_j + K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}\hat{r}_j \end{aligned} \tag{B12}$$

Note that the fitted random effect for the *j*th individual has been expressed as a linear function of the predicted random effect.

**3. Estimated and predicted errors**

Let us define $\hat{e}_j = r_j - \tilde{r}_j$ as the estimated error and $e_j = r_j - \hat{r}_j$ as the predicted error. We then define

$$\begin{aligned} r_j - \tilde{r}_j = {} & r_j - K_{jj}C_{jj}r_j + K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}r_j \\ & + K_{jj}C_{jj}\hat{r}_j - \hat{r}_j - K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}\hat{r}_j \end{aligned} \tag{B13}$$

After a few steps of manipulation, we have

$$r_j - \tilde{r}_j = \left[I - K_{jj}C_{jj} + K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}\right](r_j - \hat{r}_j) \tag{B14}$$

Therefore, the estimated and predicted errors have the following relationship,

$$\hat{e}_j = \left[I - K_{jj}C_{jj} - K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj}\right]e_j \tag{B15}$$

We want to prove that the *j*th diagonal element of the HAT matrix (the leverage value for observation *j*) is

$$H_{jj} = K_{jj}C_{jj} - K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj} \tag{B16}$$

which leads to

$$\hat{e}_j = (I - H_{jj})e_j \tag{B17}$$

As a result,

$$e_j = (I - H_{jj})^{-1}\hat{e}_j \tag{B18}$$

We now go back to the HAT matrix to see what $H_{jj}$ is. Using partitioned matrix, we have

$$\begin{aligned} H &= \begin{bmatrix} K_{jj} & K_{j(-j)} \\ K_{(-j)j} & K_{(-j)(-j)} \end{bmatrix} \begin{bmatrix} C_{jj} & C_{j(-j)} \\ C_{(-j)j} & C_{(-j)(-j)} \end{bmatrix} \\ &= \begin{bmatrix} K_{jj}C_{jj} + K_{j(-j)}C_{(-j)j} & K_{jj}C_{jj}C_{j(-j)} + K_{j(-j)}C_{(-j)(-j)} \\ K_{(-j)j}C_{jj} + K_{(-j)(-j)}C_{(-j)j} & K_{(-j)j}C_{j(-j)} + K_{(-j)(-j)}C_{(-j)(-j)} \end{bmatrix} \end{aligned} \tag{B19}$$

Therefore,

$$H_{jj} = K_{jj}C_{jj} + K_{j(-j)}C_{(-j)j} \tag{B20}$$

From equation (B11), we know

$$K_{j(-j)}C_{(-j)j} = -K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj} \tag{B21}$$

Substituting equation (B21) into equation (B20) yields

$$K_{jj}C_{jj} - K_{j(-j)}(K_{(-j)(-j)} + \lambda I)^{-1}K_{(-j)j}C_{jj} \tag{B22}$$

which is exactly the same as equation (B16) and thus we conclude the derivation of equation (B18). Unlike the HAT matrix in fixed models, the random model HAT matrix is not idempotent, although it remains symmetric.

The PRESS is now defined as

$$\text{PRESS} = \sum_{j=1}^{n} e_j^T e_j = \sum_{j=1}^{n} \hat{e}_j^T (I - H_{jj})^{-2}\hat{e}_j \tag{B23}$$

If the residual errors are heterogeneous with $e \sim N(0, R\sigma^2)$ where $R$ is a known diagonal matrix, all the above derivations apply except that we have to replace $\lambda I$ by $\lambda R$ in all occurrences. In addition, the PRESS should be modified as a weighted PRESS,

$$\text{PRESS} = \sum_{j=1}^{n} e_j^T W_j e_j = \sum_{j=1}^{n} \hat{e}_j^T (I - H_{jj})^{-1} W_j (I - H_{jj})^{-1}\hat{e}_j \tag{B24}$$

where $W_j = R_j^{-1}$ is the weight for the *j*th observation or the *j*th block of observations.

**Table 1**

Analysis of variance table to estimate heritability of agronomic traits from replicated experiments of hybrid rice

| Source | Degree of freedom | SS | MS | E(MS)[a] |
|---|---|---|---|---|
| Hybrids | 1495-1 = 1594 | $SS_G$ | $MS_G$ | $\sigma_E^2 + 2\sigma_G^2$ |
| Locations | 2-1 = 1 | $SS_R$ | $MS_R$ | $\sigma_E^2 + 1495\phi_R^2$ |
| Residual errors | (1995-1)(2-1) = 1494 | $SS_E$ | $MS_E$ | $\sigma_E^2$ |
| Corrected Total | 2989 | $SS_T$ | $MS_T$ | |

[a] These variance components are used to estimate the trait heritability $H^2 = \sigma_G^2 / (\sigma_G^2 + \sigma_E^2)$

## Figure Legends

**Figure 1.** Comparison of predictabilities of the HAT and CV methods for seven agronomy traits in inbred rice. The first seven panels are the predictabilities of the HAT (blue) and CV (red) for the seven traits. The last panel shows the average predictabilities and the 95 confidence band of the HAT method for KGW from 100 random partitionings of the sample.

**Figure 2.** Predictability of the CV method plotted against that of the HAT method under n-fold cross validation for different traits. Panel A: Seven traits of the inbred rice; Panel B: 1000 metabolomic traits of the inbred rice; Panel C: 10 traits in hybrid rice; Panel D: Plot of the $R^2$ of the CV method, the HAT method and the goodness of fit (FIT) against the estimated heritability obtained from replicated experiments for ten agronomy traits of the hybrid rice.

**Figure 3.** Comparison of $R^2$ of the estimated heritability from replicated experiments (H2), the CV method, the HAT method and the model goodness of fit (FIT) for ten traits of the hybrid rice.

**Figure 4.** Tuning parameter ( $\lambda = \sigma^2 / \sigma_\xi^2$ ) that maximizes the genomic predictability ( $R^2$ ) of human height. Panel A: predictability profile of the HAT method, where the red point represents the predictability ( $R^2_{\text{HAT}} = 0.3278$ ) when the tuning parameter takes the REML estimate ( $\hat{\lambda} = 0.1366$ ) and the blue point represents the maximum achievable predictability ( $R^2_{\text{HAT}} = 0.3310$ ) when the tuning parameter is $\lambda = 0.2180$. Panel B: predictability profile of the GCV method, where the red point represents the predictability ( $R^2_{\text{GCV}} = 0.3536$ ) when the tuning parameter takes the REML estimate ( $\hat{\lambda} = 0.1366$ ) and the blue point represents the maximum predictability ( $R^2_{\text{GCV}} = 0.3539$ ) when the tuning parameter is $\lambda = 0.1580$.

**Figure 5.** Comparison of predictability of REML estimated $\lambda$ with the maximum achievable predictability by tuning $\lambda$ for 1000 metabolomic traits of rice. Panel A: maximum achievable predictability of the HAT method by tuning $\lambda$ plotted against the predictability when $\lambda$ takes the REML estimate. Panel B: maximum achievable predictability of the GCV method by tuning $\lambda$ plotted against the predictability when $\lambda$ takes the REML estimate. Panel C: predictability of the GCV method plotted against the predictability of the HAT method when $\lambda$ takes the REML estimate. The red points indicate traits with visible differences in predictability between the method shown in the x-axis and the method shown in the y-axis. The linear regression equation is given at the bottom of each panel and the fitted $r^2$ of the regression is given at the top of each panel.

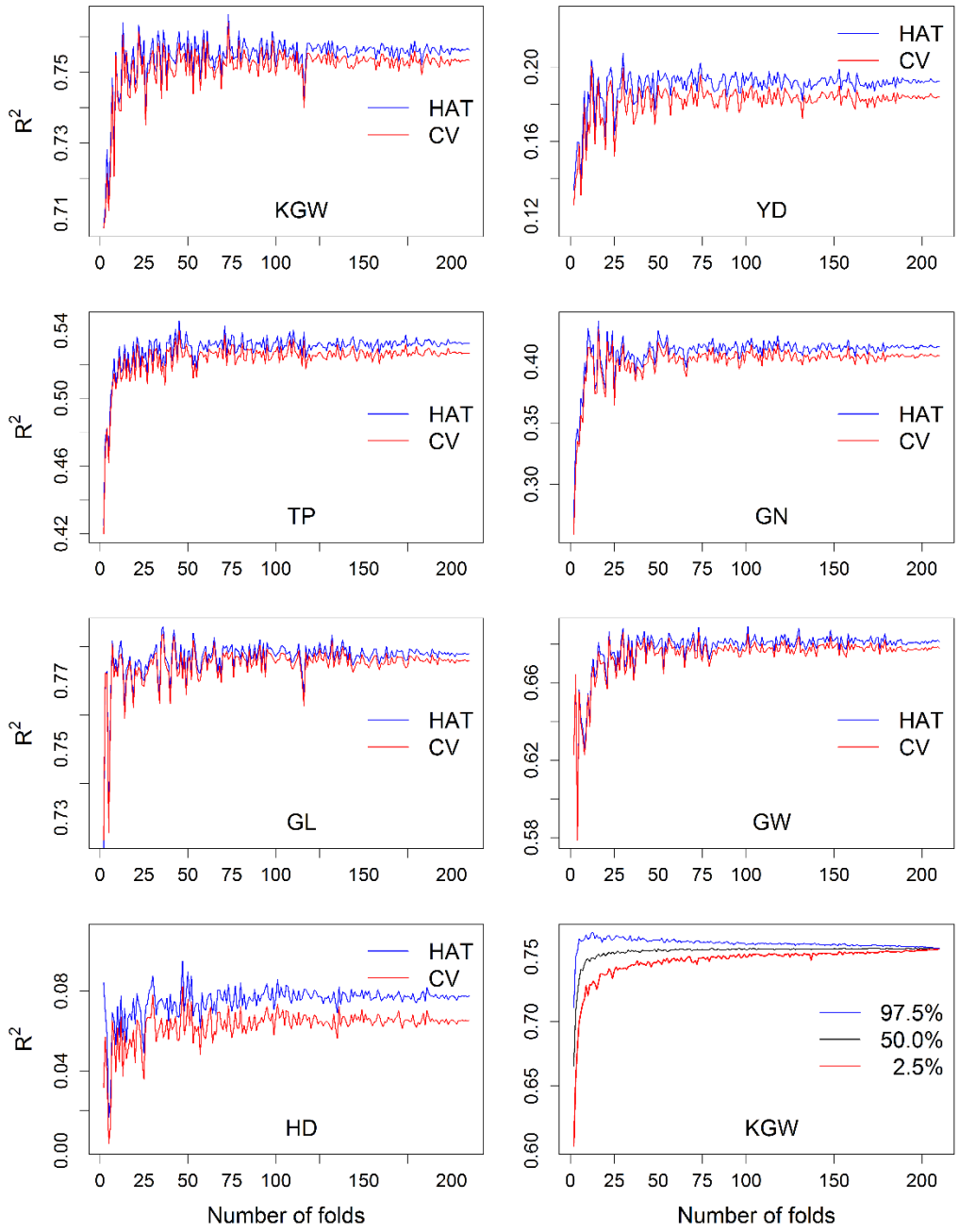


**Figure 1.** Comparison of predictabilities of the HAT and CV methods for seven agronomy traits in inbred rice. The first seven panels are the predictabilities of the HAT (blue) and CV (red) for the seven traits. The last panel shows the average predictabilities and the 95 confidence band of the HAT method for KGW from 100 random partitionings of the sample.

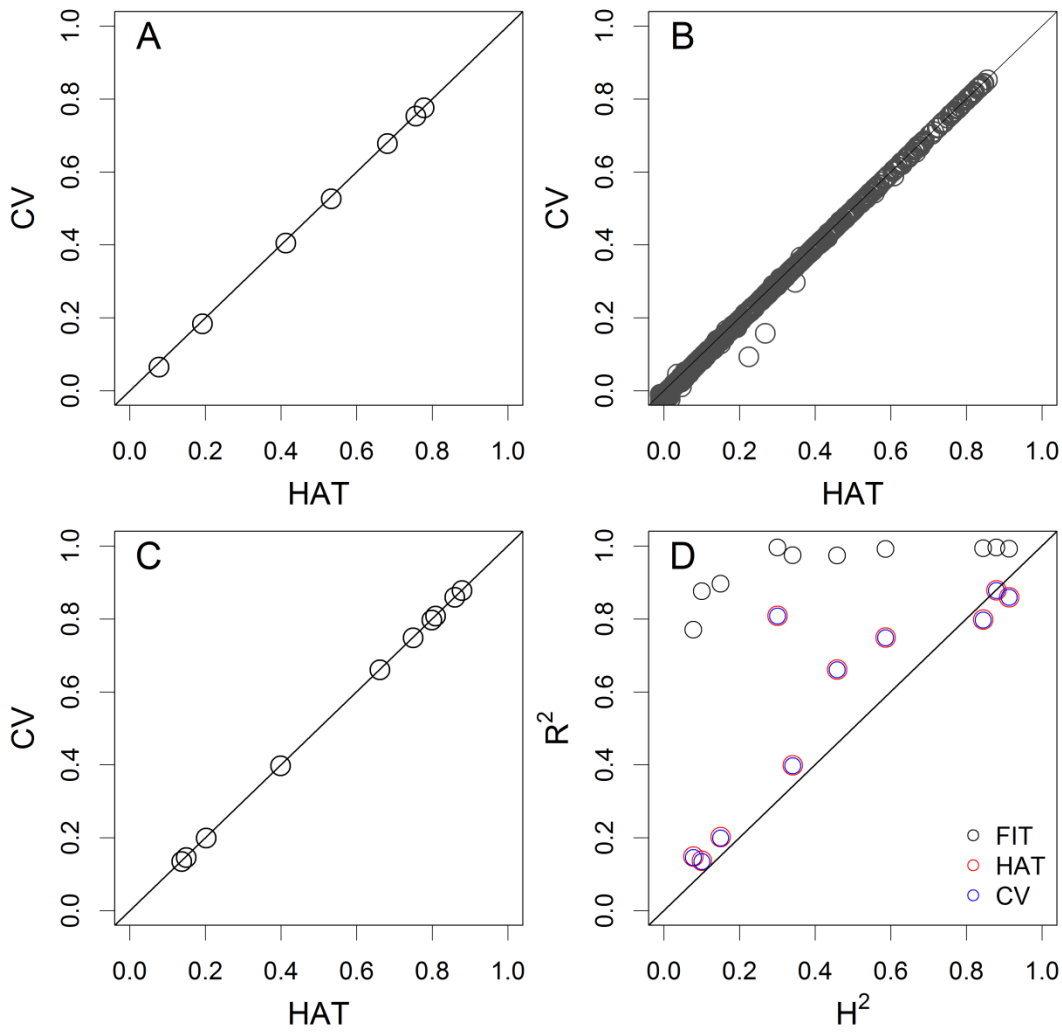


**Figure 2.** Predictability of the CV method plotted against that of the HAT method under n-fold cross validation for different traits. Panel A: the QQ-plot of the seven traits of the inbred rice; Panel B: the QQ-plot of 1000 metabolomic traits of the inbred rice; Panel C: the QQ-plot of 10 traits in hybrid rice; Panel D: Plot of the $R^2$ of the CV method, the HAT method and the goodness of fit (FIT) against the estimated heritability obtained from replicated experiments for ten agronomy traits of the hybrid rice.

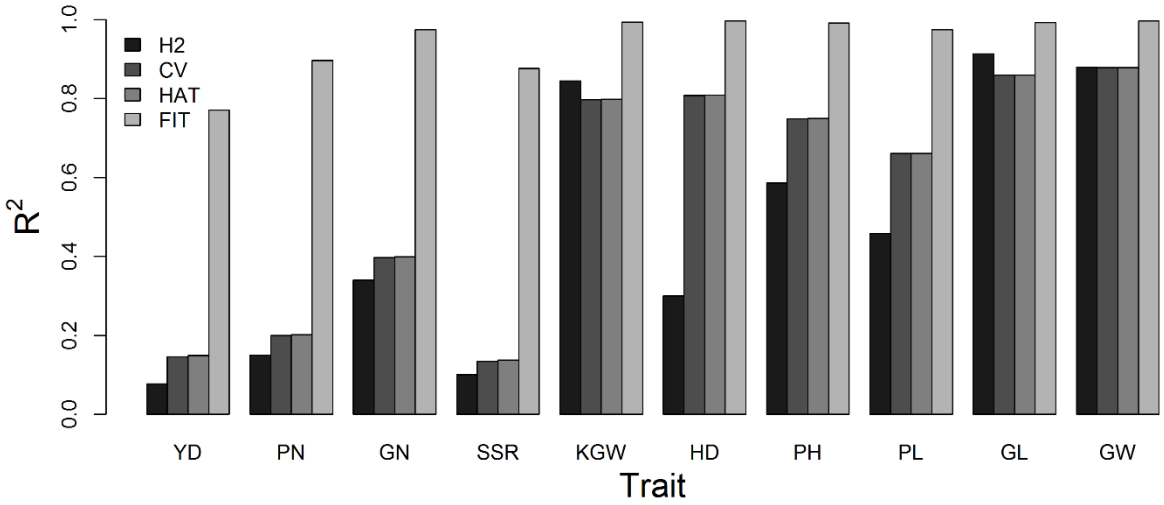


**Figure 3.** Comparison of $R^2$ of the estimated heritability from replicated experiments (H2), the CV method, the HAT method and the model goodness of fit (FIT) for ten traits of the hybrid rice.

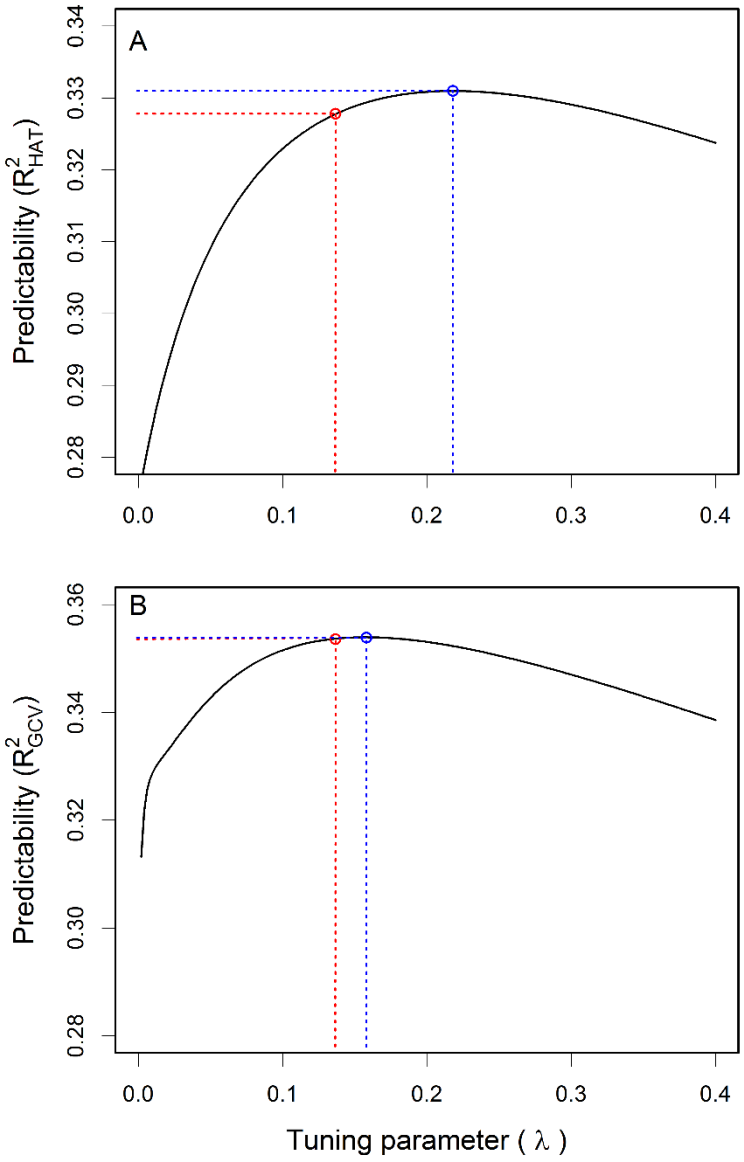


**Figure 4.** Tuning parameter ( $\lambda = \sigma^2 / \sigma^2_\xi$ ) that maximizes the genomic predictability ( $R^2$ ) of human height. Panel A: predictability profile of the HAT method, where the red point represents the predictability ( $R^2_{\text{HAT}} = 0.3278$ ) when the tuning parameter takes the REML estimate ( $\hat{\lambda} = 0.1366$ ) and the blue point represents the maximum achievable predictability ( $R^2_{\text{HAT}} = 0.3310$ ) when the tuning parameter is $\lambda = 0.2180$. Panel B: predictability profile of the GCV method, where the red point represents the predictability ( $R^2_{\text{GCV}} = 0.3536$ ) when the tuning parameter takes the REML estimate ( $\hat{\lambda} = 0.1366$ ) and the blue point represents the maximum predictability ( $R^2_{\text{GCV}} = 0.3539$ ) when the tuning parameter is $\lambda = 0.1580$.

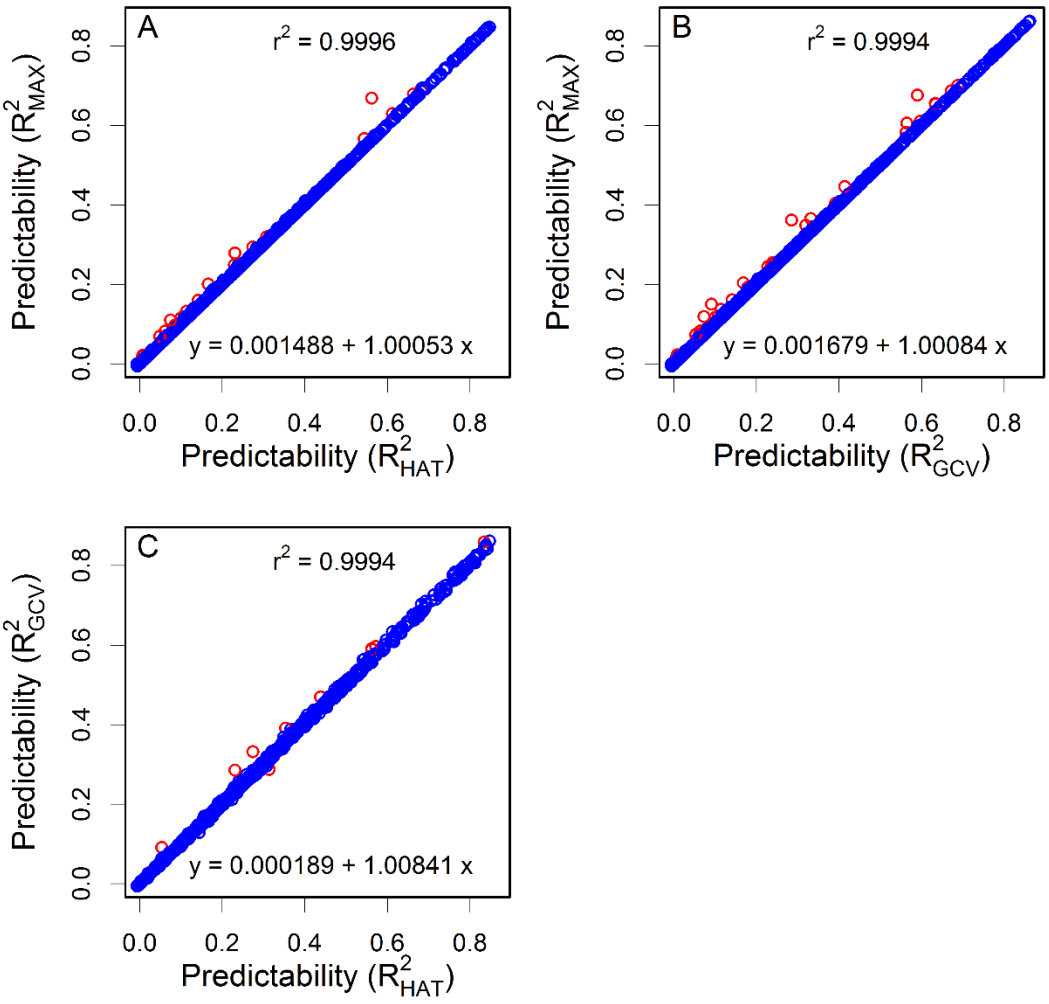


**Figure 5.** Comparison of predictability of REML estimated $\lambda$ with the maximum achievable predictability by tuning $\lambda$ for 1000 metabolomic traits of rice. Panel A: maximum achievable predictability of the HAT method by tuning $\lambda$ plotted against the predictability when $\lambda$ takes the REML estimate. Panel B: maximum achievable predictability of the GCV method by tuning $\lambda$ plotted against the predictability when $\lambda$ takes the REML estimate. Panel C: predictability of the GCV method plotted against the predictability of the HAT method when $\lambda$ takes the REML estimate. The red points indicate traits with visible differences in predictability between the method shown in the x-axis and the method shown in the y-axis. The linear regression equation is given at the bottom of each panel and the fitted $r^2$ of the regression is given at the top of each panel.