Cheng *et al. Journal of Animal Science and Biotechnology* (2017) 8:38
DOI 10.1186/s40104-017-0164-6

Journal of Animal Science and Biotechnology

METHODOLOGY ARTICLE Open Access

Efficient strategies for leave-one-out cross validation for genomic best linear unbiased prediction

Hao Cheng[1,2], Dorian J. Garrick[1,3] and Rohan L. Fernando[1*]

Abstract

Background: A random multiple-regression model that simultaneously fit all allele substitution effects for additive markers or haplotypes as uncorrelated random effects was proposed for Best Linear Unbiased Prediction, using whole-genome data. Leave-one-out cross validation can be used to quantify the predictive ability of a statistical model.

Methods: Naive application of Leave-one-out cross validation is computationally intensive because the training and validation analyses need to be repeated n times, once for each observation. Efficient Leave-one-out cross validation strategies are presented here, requiring little more effort than a single analysis.

Results: Efficient Leave-one-out cross validation strategies is 786 times faster than the naive application for a simulated dataset with 1,000 observations and 10,000 markers and 99 times faster with 1,000 observations and 100 markers. These efficiencies relative to the naive approach using the same model will increase with increases in the number of observations.

Conclusions: Efficient Leave-one-out cross validation strategies are presented here, requiring little more effort than a single analysis.

Keywords: Leave-one-out cross validation, GBLUP

Background

A random multiple-regression model that simultaneously fit all allele substitution effects for additive markers or haplotypes as uncorrelated random effects was proposed for Best Linear Unbiased Prediction (BLUP) [1], using whole-genome data. Breeding values are defined as the sum of the effects of all the markers or haplotypes, and their estimates are widely used for prediction of the merit of selection candidates. Estimates of marker or haplotype effects are used to predict breeding values of individuals that were not present in a previous analysis commonly referred to as training. An alternative earlier published approach to use marker or haplotype information fits breeding values as random effects based on covariances defined by a "genomic relationship matrix" computed from genotypes [2]. These two models have been shown to be equivalent in terms of predicting breeding values [3, 4] and we refer to them here as marker effect models (MEM) or breeding value models (BVM), the latter often known as Genomic Best Linear Unbiased Prediction (GBLUP).

Cross validation is often used to quantify the predictive ability of a statistical model. In k-fold cross validation, the whole dataset is partitioned into k parts with k analyses, where one part is omitted for training with validation on the omitted part. Leave-one-out cross validation (LOOCV) is a special case of k-fold cross validation with $k = n$, the number of observations. When the dataset is small, leave-one-out cross validation is appealing as the size of the training set is maximized. However, naive application of LOOCV is computationally intensive, requiring n analyses.

We show below how LOOCV can be performed using either the MEM or BVM with little more effort than is required for a single analysis with n observations.

Methods

Use of the MEM is more efficient when the number n of individuals is larger than the number p of markers,

*Correspondence: rohan@iastate.edu
[1]Department of Animal Science, Iowa State University, 50011, Ames, Iowa, USA
Full list of author information is available at the end of the article

because for this model the mixed model equations are of order p plus the number of other effects. When $n < p$, estimated breeding values can be obtained more efficiently by solving the mixed model equations for the BVM of order n plus the number of other effects. We deal with the special case where the only other effect is a general mean and phenotypes have been pre-corrected for other nuisance variables. Efficient strategies for LOOCV using this special case for MEM when $n \geq p$ and BVM when $p \geq n$ are shown below.

Marker effect models

The MEM for GBLUP can be written as

$$\boldsymbol{y} = \mathbf{1}\mu + \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{e}, \tag{1}$$

where $\boldsymbol{y}$, a $n \times 1$ vector for phenotypes, has been pre-corrected for all fixed effects other than μ, the overall mean, $\boldsymbol{X}$ is the $n \times p$ matrix of marker covariates, $\boldsymbol{\beta}$ is a $p \times 1$ random vector of the allele substitution effects and $\boldsymbol{e}$ is a $n \times 1$ random vector of residuals. Often it is assumed that marker effects are identically and independently distributed (iid) random variables with null means and variances σ^2_β. Thus, under the usual assumption that the residuals are iid with null means and variances σ^2_e, $E(\mathbf{y}) = \mathbf{1}\mu$. When MEM is used, LOOCV can be performed by using a well-known strategy used in least-squares regression to compute the predicted residual sum of square (PRESS) [5] statistic.

LOOCV strategy for MEM

BLUP of $\boldsymbol{\beta}^* = \begin{bmatrix} \mu \\ \boldsymbol{\beta} \end{bmatrix}$ can be obtained by solving the mixed model equations

$$\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)\hat{\boldsymbol{\beta}}^* = \boldsymbol{X}^{*\prime}y, \tag{2}$$

where $\boldsymbol{X}^* = [\mathbf{1} \;\; \boldsymbol{X}]$, $\hat{\boldsymbol{\beta}}^* = \begin{bmatrix} \hat{\mu} \\ \hat{\boldsymbol{\beta}} \end{bmatrix}$, $\boldsymbol{D}$ is a diagonal matrix whose elements are 0 followed by a p vector of 1s and $\lambda = \frac{\sigma^2_e}{\sigma^2_\beta}$.

Now, BLUP for $\boldsymbol{\beta}^*_{-j}$, where observation j is left out, can be obtained as

$$\hat{\boldsymbol{\beta}}^*_{-j} = \left(\boldsymbol{X}^{*\prime}_{-j}\boldsymbol{X}^*_{-j} + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{X}^{*\prime}{}_{-j}\boldsymbol{y}_{-j}, \tag{3}$$

where $\boldsymbol{X}^*_{-j}$ is $\boldsymbol{X}^*$ with the jth row removed and $\boldsymbol{y}_{-j}$ is $\boldsymbol{y}$ with the jth element removed.

Suppose $\boldsymbol{x}^{*\prime}_j$ is the jth row of $\boldsymbol{X}^*$, then from the matrix inverse lemma [4],

$$\begin{aligned}\left(\boldsymbol{X}^{*\prime}_{-j}\boldsymbol{X}^*_{-j} + \boldsymbol{D}\lambda\right)^{-1} &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda - \boldsymbol{x}^*_j\boldsymbol{x}^{*\prime}_j\right)^{-1} \\ &= \left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1} \\ &\quad - \frac{\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{x}^*_j\boldsymbol{x}^{*\prime}_j\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}}{1 - H_{jj}},\end{aligned} \tag{4}$$

where the quadratic $H_{jj} = \boldsymbol{x}^*_j\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{x}^*_j$ is the jth diagonal element of $\boldsymbol{H} = \boldsymbol{X}^*\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{X}^{*\prime}$.

Using (3) in (4), the prediction residual for the jth observation can be written as

$$\begin{aligned}\hat{e}_j &= y_j - \boldsymbol{x}^{*\prime}_j\hat{\boldsymbol{\beta}}^*{}_{-j} \\ &= y_j - \boldsymbol{x}^{*\prime}_j\left[\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\right. \\ &\quad \left. - \frac{\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{x}^*_j\boldsymbol{x}^{*\prime}_j\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}}{1 - H_{jj}}\right]\boldsymbol{X}^{*\prime}_{-j}\boldsymbol{y}_{-j} \\ &= \frac{\left(1 - \boldsymbol{H}_{jj}\right)y_j - \boldsymbol{x}^{*\prime}_j\left(\boldsymbol{X}^{*\prime}\boldsymbol{X}^* + \boldsymbol{D}\lambda\right)^{-1}\boldsymbol{X}^*_{-j}\boldsymbol{y}_{-j}}{1 - \boldsymbol{H}_{jj}} \qquad (5) \\ &= \frac{y_j - \boldsymbol{x}^{*\prime}_j\hat{\boldsymbol{\beta}}^*}{1 - \boldsymbol{H}_{jj}}. \qquad (6)\end{aligned}$$

These prediction errors can be squared and accumulated over n realizations to compute PRESS defined as $\sum_{j=1}^{n}\hat{e}_j{}^2$. The accuracy of genomic prediction is often quantified as the correlation between the predicted and observed values of y_j, and that correlation can be estimated from the values of $\hat{y}_j$, which can be computed efficiently as $\hat{y}_j = y_j - \hat{e}_j$, using the observed values of y_j. When a specific group of individuals is of interest, prediction accuracies and PRESS can also be calculate using $\hat{e}_j$ for individuals in that group.

Breeding value models

When $n < p$, the genomic prediction of the breeding value $\mathbf{x}'_j\hat{\boldsymbol{\beta}}$ can be obtained more efficiently by solving the mixed model equations for the BVM:

$$\boldsymbol{y} = \mathbf{1}\mu + \mathbf{Z}\mathbf{u} + \boldsymbol{e}, \tag{7}$$

where $\mathbf{u} = \mathbf{X}\beta$, $\mathrm{var}(\mathbf{u}) = \mathbf{X}\mathbf{X}'\sigma^2_\beta$, $\mathbf{Z}$ is the identity matrix of order n and other variables are as in the MEM. Further, in both models $\mathrm{E}(\mathbf{y}) = \mathbf{1}\mu$, and $\mathrm{var}(\mathbf{y}) = \mathbf{X}\mathbf{X}'\sigma^2_\beta + \mathbf{I}\sigma^2_e$. These two models are said to be equivalent [6], and linear functions predicted from one model are identical to corresponding predictions from the other model. Two efficient strategies for LOOCV using the BVM are shown below.

LOOCV strategy I for BVM

The mixed model equations for this model are:

$$\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*}+\boldsymbol{G}\lambda\right)\hat{\boldsymbol{u}}^{*}=\boldsymbol{Z}^{*\prime}y, \tag{8}$$

where $\boldsymbol{Z}^{*}=[\mathbf{1}\quad \boldsymbol{Z}]$, $\hat{\boldsymbol{u}}^{*}=\begin{bmatrix}\hat{\mu}\\ \hat{\boldsymbol{u}}\end{bmatrix}$, $\mathbf{G}=\begin{bmatrix}0 & \mathbf{0}^{'}\\ \mathbf{0} & \left(XX'\right)^{-1}\end{bmatrix}$. Due to the relative order of the coefficient matrices for the MEM and the BVM, when $n < p$, $\boldsymbol{x}_j^{*\prime}\hat{\boldsymbol{\beta}}^{*}$ is more efficiently obtained as $\hat{u^{*}}_j$. Similarly, $\text{var}(\boldsymbol{x}_j^{*\prime}\boldsymbol{\beta}^{*}-\boldsymbol{x}_j^{*\prime}\hat{\boldsymbol{\beta}}^{*})=\mathbf{x}^{*\prime}{}_j\left(\mathbf{X}^{*\prime}\mathbf{X}^{*}+\mathbf{D}\lambda\right)^{-1}\mathbf{x}_j^{*}\sigma_e^2$ can be obtained more efficiently as $\text{var}(u_j^{*}-\hat{u^{*}}_j)=\mathbf{z}^{*}{}_{j'}\left(\mathbf{Z}^{*\prime}\mathbf{Z}^{*}+\mathbf{G}\lambda\right)^{-1}\mathbf{z}^{*}{}_j\sigma_e^2$. Using these two equalities, the formula for $\hat{e}_j$ becomes:

$$\begin{aligned}\hat{e}_j &= y_j-\boldsymbol{x}_j^{*\prime}\hat{\boldsymbol{\beta}}^{*}{}_{-j}\\ &=\frac{y_j-\boldsymbol{x}_j^{*\prime}\hat{\boldsymbol{\beta}}^{*}}{1-H_{jj}}\\ &=\frac{y_j-\boldsymbol{x}_j^{*\prime}\hat{\boldsymbol{\beta}}^{*}}{1-\boldsymbol{x}_j^{*\prime}\left(\mathbf{X}^{*\prime}\mathbf{X}^{*}+\mathbf{D}\lambda\right)^{-1}\mathbf{x}_j^{*}}\\ &=\frac{y_j-\mathbf{z}_j^{*\prime}\hat{u^{*}}}{1-\mathbf{z}_j^{*\prime}\left(\mathbf{Z}^{*\prime}\mathbf{Z}^{*}+\mathbf{G}\lambda\right)^{-1}\mathbf{z}_j^{*}}\\ &=\frac{y_j-\mathbf{z}_j^{*\prime}\hat{u^{*}}}{1-\boldsymbol{C}_{jj}},\end{aligned} \tag{9}$$

where the quadratic $\mathbf{z}_j^{*\prime}\left(\mathbf{Z}^{*\prime}\mathbf{Z}^{*}+\mathbf{G}\lambda\right)^{-1}\mathbf{z}_j^{*}$ is the jth diagonal element of $\boldsymbol{C}=\boldsymbol{Z}^{*}\left(\boldsymbol{Z}^{*\prime}\boldsymbol{Z}^{*}+\boldsymbol{G}\lambda\right)^{-1}\boldsymbol{Z}^{*\prime}$.

LOOCV strategy II for BVM

Another efficient strategy for BVM is shown here. First we consider the situation where $\boldsymbol{y}$ has been pre-corrected for μ in addition to nuisance effects so that $\text{E}(\mathbf{y})=\mathbf{0}$ and we define $\mathbf{var}\left(\boldsymbol{y}\right)=\mathbf{X}\mathbf{X}'\sigma_\beta^2+\mathbf{I}\sigma_e^2=V$. Now matrix $\mathbf{Q}$ is constructed by augmenting the covariance matrix of $\boldsymbol{y}$ with one leading row and column as

$$\mathbf{Q}=\begin{bmatrix}\mathbf{y}'\mathbf{y} & \mathbf{y}'\\ \mathbf{y} & \mathbf{V}\end{bmatrix}.$$

To obtain the prediction error for observation j, the second row and column of $\mathbf{Q}$ are permuted with row and column $j+1$. In this manner $\mathbf{Q}$ has its rows and columns symmetrically permuted as $\mathbf{P}^{'}{}_j\mathbf{Q}\mathbf{P}_j=\boldsymbol{W}$, where the permutation matrix $\mathbf{P}_j$ is obtained by permuting the second row of the n order identity matrix with row $j+1$. So the permuted matrix is:

$$\mathbf{W}=\begin{bmatrix}\mathbf{y}'\mathbf{y} & y_j & \mathbf{y}'_{-j}\\ y_j & V_{jj} & \mathbf{V}_{j,-j}\\ \mathbf{y}_{-j} & \mathbf{V}_{-j,j} & \mathbf{V}_{-j,-j}\end{bmatrix}=\begin{bmatrix}\mathbf{A} & \mathbf{B}\\ \mathbf{B}^{'} & \mathbf{C}\end{bmatrix}$$

where we will define the leading 2×2 matrix as $\mathbf{A}=\begin{bmatrix}\mathbf{y}'\mathbf{y} & y_j\\ y_j & V_{jj}\end{bmatrix}$, and the other partitions as $\mathbf{B}=\begin{bmatrix}\mathbf{y}'_{-j}\\ \mathbf{V}_{j,-j}\end{bmatrix}$, and $\mathbf{C}=\mathbf{V}_{-j,-j}$, where $-j$ denotes that the jth element, row or column has been removed. Defining $\mathbf{W^{11}}$ as the top left or leading 2×2 sub-matrix in $\mathbf{W^{-1}}$ corresponding to the position of $\mathbf{A}$ in $\mathbf{W}$, and using partitioned inverse-matrix identities [7], the inverse of $\mathbf{W^{11}}$ can be written as,

$$\begin{aligned}\left(\mathbf{W^{11}}\right)^{-1} &= \mathbf{A}-\mathbf{B}\mathbf{C^{-1}}\mathbf{B}'\\ &=\begin{bmatrix}\mathbf{y}'\mathbf{y} & y_j\\ y_j & V_{jj}\end{bmatrix}-\begin{bmatrix}\mathbf{y}'_{-\mathbf{j}}\\ \mathbf{V}_{\mathbf{j},-\mathbf{j}}\end{bmatrix}\mathbf{V}^{-1}_{-j,-j}\begin{bmatrix}\mathbf{y}_{-\mathbf{j}} & \mathbf{V}_{-\mathbf{j},\mathbf{j}}\end{bmatrix}\\ &=\begin{bmatrix}\mathbf{y}'\mathbf{y}-\mathbf{y}'_{-j}\mathbf{V}^{-\mathbf{1}}_{-\mathbf{j},-\mathbf{j}}\mathbf{y}_{-j} & y_j-\mathbf{y}'_{-j}V^{-1}_{-j,-j}V_{-j,j}\\ y_j-\mathbf{V}_{\mathbf{j},-\mathbf{j}}\mathbf{V}^{-\mathbf{1}}_{-\mathbf{j},-\mathbf{j}}\mathbf{y}_{-\mathbf{j}} & V_{jj}-\mathbf{V}_{\mathbf{j},-\mathbf{j}}\mathbf{V}^{-\mathbf{1}}_{-\mathbf{j},-\mathbf{j}}\mathbf{V}_{-\mathbf{j},\mathbf{j}}\end{bmatrix}.\end{aligned} \tag{10}$$

Now $\boldsymbol{V_{j,-j}}$ in element $(2,1)$ of the above inverse matrix is the vector of covariances between y_j and $\mathbf{y}_{-j}$ and $\mathbf{V}^{-\mathbf{1}}_{-\mathbf{j},-\mathbf{j}}$ is the inverse of the covariance matrix of $\mathbf{y}_{-j}$. Thus, $\hat{y}_j=V_{j,-j}V^{-1}_{-j,-j}\mathbf{y}_{-j}$ is the Best Linear Predictor (BLP) of y_j given $\mathbf{y}_{-j}$, and element (2,1) of (10) is the prediction error of y_j. The element $(2,2)$ in (10) is the prediction error variance (PEV) for y_j, where $PEV=var(y_j-\hat{y}_j)$. PEV can also be used to calculate theoretical reliability for individual i as $1-\frac{PEV_i}{V_{jj}}$, and characterizing the distributions of reliability for all the individuals in a dataset has a number of practical applications. Note this allows us to obtain the PEV of every individual and the distribution of these values provide information as to the robustness of genomic predictions across the population of individuals represented in the dataset. This PEV is determined by the genomic variance-covariance matrix and does not depend on $\mathbf{y}$. Two different datasets could generate the same PRESS statistic but with different distributions of PEV.

Now, because the permutation matrix $\mathbf{P}_j$ is orthogonal, $\mathbf{W}^{-1}=(\mathbf{P}'_j\mathbf{Q}\mathbf{P}_j)^{-1}=\mathbf{P}'_j\mathbf{Q}^{-1}\mathbf{P}_j$, and the elements of $\mathbf{W}^{11}$ that are of interest in terms of predicting individual j can be obtained directly from $\mathbf{Q}^{-1}$ as

$$\mathbf{W}^{11}=\begin{bmatrix}q^{1,1} & q^{1,(1+j)}\\ q^{(1+j),1} & q^{(1+j),(1+j)}\end{bmatrix}. \tag{11}$$

It follows that $\hat{e}_j$, which is the off-diagonal element of the inverse of the 2×2 matrix $\mathbf{W}^{11}$, can be written in terms of $\mathbf{Q}^{-1}$ as

$$\hat{e}_j=\frac{-q^{(1+j),1}}{q^{1,1}q^{(1+j),(1+j)}-q^{1,(1+j)}q^{(1+j),1}}, \tag{12}$$

where $q^{i,j}$ is the element from row i and column j of $\mathbf{Q}^{-1}$. Thus, once $\mathbf{Q}^{-1}$ is computed, $\hat{e}_j$ for all j can be com-

Table 1 Phenotypes and genotypes at 5 markers for 3 individuals used in the numerical example

	M1	M2	M3	M4	M5	Phenotypes
1	1	2	1	2	2	1.97
2	2	1	0	1	1	2.12
3	0	0	2	1	2	–0.62

puted using (12), and these values can be used to compute PRESS as $\sum_{j=1}^{n} \hat{e}_j^2$. To estimate the correlation between the predicted and observed values of y_j, the value of $\hat{y}_j$ is efficiently obtained as the difference $\hat{y}_j = y_j - \hat{e}_j$.

Now we consider the situation without pre-correcting $\mathbf{y}$ for μ, where $\mathrm{E}(\mathbf{y}) = \mathbf{1}\mu$. Now the mixed model (7) contains both fixed and random effects. Note that the mixed model equations that correspond to this mixed effects model can be derived by treating μ as "random" with null mean and large variance. So, let

$$var\left(\boldsymbol{\beta}^*\right) = \begin{bmatrix} \sigma_L^2 & \mathbf{0}' \\ \mathbf{0} & I\sigma_\beta^2 \end{bmatrix} = \boldsymbol{\Sigma},$$

for sufficiently large value of σ_L^2. Then under this assumption, $\mathrm{E}(\mathbf{y}) = \mathbf{0}$ and $var\left(\boldsymbol{y}\right) = \mathbf{X}^*\boldsymbol{\Sigma}\boldsymbol{X}^{*'} + \mathbf{I}\sigma_e^2 = \boldsymbol{V}^*$, and thus $\hat{y}_j = \boldsymbol{V}_{j,-j}^* \boldsymbol{V}_{-j,-j}^{*-1} \mathbf{y}_{-j}$ is the BLP from the random effects rather than mixed effects model of y_j given $\mathbf{y}_{-j}$. This BLP obtained from the model with random μ will be numerically very close to the BLUP obtained from the mixed model with fixed μ. The $\mathbf{Q}$ matrix corresponding to the BLP with random μ is constructed as $\begin{bmatrix} \mathbf{y}'\mathbf{y} & \mathbf{y}' \\ \mathbf{y} & \mathbf{V}^* \end{bmatrix}$ and prediction residuals are obtained as (12).

Numerical example

Phenotypes $\boldsymbol{y}$ and genotypes $\boldsymbol{X}$ at 5 markers for 3 individuals are in Table 1. Assume $\sigma_\beta^2 = \frac{\sigma_e^2}{10}$ and the overall mean μ is the only fixed effect. In LOOCV strategy for MEM and strategy I for BVM, the diagonal elements of $\boldsymbol{H}$ for MEM and $\boldsymbol{C}$ for BVM, which are in the denominators of (6) and (9), are in Table 2. The numerators of (6) and (9) are obtained by solving the MME (2) and (8). Then prediction errors are calculated as in (6) and (9) and shown in Table 4. In LOOCV strategy II for BVM, the $\boldsymbol{Q}$ matrix (Table 3) is constructed using $\sigma_L^2 = 1000$, which is sufficiently large relative to σ_e^2 for μ to be indistinguishable from a fixed effect with a flat prior. The prediction errors are calculated as (12) and shown in Table 4. The MEM strategy and BVM strategy I gave identical prediction errors and identical PRESS for this numerical example were numerically very close to those from the BVM strategy II.

Table 2 Diagonal elements of $\boldsymbol{H}$ in LOOCV strategy for MEM and $\boldsymbol{C}$ for BVM

	$j = 1$	$j = 2$	$j = 3$
H_{jj}	0.46	0.51	0.55
C_{jj}	0.46	0.51	0.55

Table 3 Q matrix in strategy II for BVM

	1	2	3	4
1	8.75	1.97	2.12	–0.62
2	1.97	1,002.40	1,000.80	1,000.80
3	2.12	1,000.80	1,001.70	1,000.30
4	–0.62	1,000.80	1,000.30	1,001.90

Simulation to compare efficiency

Two datasets were simulated using XSim [8], where 1,000 offspring were sampled from random mating of 100 parents for 10 non-overlapping generations, to compare the computational efficiencies for naive and efficient strategies using BVM or MEM for LOOCV in GBLUP. Dataset I was simulated with 1,000 observations and 10,000 SNP markers for a $p \gg n$ scenario. Dataset II was simulated with 1,000 observations and 100 markers for a $n \gg p$ scenario. The processor used in the analyses was a 1.4 GHz Intel Core i5 with 4 GB of memory.

For dataset II, efficient MEM is 99 times faster than the naive application (2.979 s versus 0.030 s) (Table 5). All strategies implemented in Julia, a scientific programming language, gave virtually identical prediction accuracies defined as the correlation between $\mathbf{y}$ and $\hat{\mathbf{y}}$ for each dataset. For dataset I, efficient BVM is 786 times faster than the naive application (3.107 s versus 2,442.59 s) (Table 5).

Discussion

In genomic prediction, the candidates to be predicted are often offspring that are genotyped but not yet phenotyped. In this situation, LOOCV using all individuals in the training dataset will provide an upper bound for the accuracy of prediction, because ancestors in the training dataset with large numbers of descendants have more accurate predictions than descendants. A better estimate of the accuracy of prediction can be obtained by applying LOOCV to only terminal offspring in the training dataset.

Table 4 Prediction errors from different LOOCV strategies (different strategies gave identical prediction errors)

	$j = 1$	$j = 2$	$j = 3$
$\hat{e}_j$	1.13	1.21	–2.66

Table 5 Efficiency of alternative LOOCV strategies for GBLUP

	Alternative LOOCV strategies				
	Naive MEM	Naive BVM	Efficient MEM	Efficient BVM I	Efficient BVM II
$n = 1,000; p = 10,000$	9,490.608	2,442.590	105.141	3.107	5.945
$n = 1,000; p = 100$	2.979	169.928	0.030	2.725	0.217

Results are given for the computing time in seconds using naive MEM, naive BVM, efficient MEM, efficient BVM I and efficient BVM II

Conclusions

Efficient strategies for LOOCV in GBLUP are presented in this paper. LOOCV strategy I and II for BVM are more efficient when $p \gg n$. LOOCV strategy for MEM is more efficient when $n \gg p$. The accuracy of genomic prediction is often quantified as the correlation between the predicted and observed values of y_j, and this correlation can be estimated efficiently using LOOCV strategies. Compared to naive application of LOOCV, which is computationally intensive, LOOCV can be implemented efficiently.

Acknowledgements
Not applicable.

Funding
This work was supported by the US Department of Agriculture, Agriculture and Food Research Initiative National Institute of Food and Agriculture Competitive grant no. 2015-67015-22947. The funders had no role in the design of the study and collection, analysis, and interpretation of data and in writing of the manuscript.

Availability of data and materials
The datasets during and/or analysed during the current study available from the corresponding author on reasonable request.

Authors' contributions
All authors contributed to the development of the statistical methods. HC wrote the program code and conducted the analyses. The manuscript was prepared by HC, RLF and DG. All authors read and approved the final manuscript.

Competing interests
The authors declare that they have no competing interests.

Consent for publication
Not applicable.

Ethics approval and consent to participate
Not applicable.

Author details
[1]Department of Animal Science, Iowa State University, 50011, Ames, Iowa, USA. [2]Department of Statistics, Iowa State University, 50011, Ames, Iowa, USA. [3]Institute of Veterinary, Animal & Biomedical Science, Massey University, Palmerston North, New Zealand.

Received: 21 September 2016 Accepted: 27 March 2017
Published online: 02 May 2017

References

1. Meuwissen THE, Hayes BJ, Goddard ME. Prediction of total genetic value using genome-wide dense marker maps. Genetics. 2001;157:1819–29.
2. Nejati-Javaremi A, Smith C, Gibson JP. Effect of total allelic relationship on accuracy of evaluation and response to selection. J Anim Sci. 1997;75:1738–45.
3. Fernando RL. Genetic evaluation and selection using genotypic, phenotypic and pedigree information. Proc 6th Wld Cong Genet Appl Livest Prod. 1998;26:329–36.
4. Strandén I, Garrick DJ. Technical note: Derivation of equivalent computing algorithms for genomic predictions and reliabilities of animal merit. J Dairy Sci. 2009;92(6):2971–75.
5. Allen DM. The relationship between variable selection and data augmentation and a method for prediction. Technometrics. 1974;16(1):125–7.
6. Henderson CR. Applications of Linear Models in Animal Breeding. Guelph: Univ. Guelph; 1984.
7. Searle SR. Matrix Algebra Useful for Statistics. New York: John Wildy and Sons, Inc; 1982.
8. Cheng H, Garrick D, Fernando R. Xsim: Simulation of descendants from ancestors with sequence data. G3. 2015;5(7):1415–17.