BMC Bioinformatics

Open Access

CrossMark

Genomic prediction using subsampling

Alencar Xavier[1], Shizhong Xu[2], William Muir[3] and Katy Martin Rainey[1*]

Abstract

Background: Genome-wide assisted selection is a critical tool for the genetic improvement of plants and animals. Whole-genome regression models in Bayesian framework represent the main family of prediction methods. Fitting such models with a large number of observations involves a prohibitive computational burden. We propose the use of subsampling bootstrap Markov chain in genomic prediction. Such method consists of fitting whole-genome regression models by subsampling observations in each round of a Markov Chain Monte Carlo. We evaluated the effect of subsampling bootstrap on prediction and computational parameters.

Results: Across datasets, we observed an optimal subsampling proportion of observations around 50% with replacement, and around 33% without replacement. Subsampling provided a substantial decrease in computation time, reducing the time to fit the model by half. On average, losses on predictive properties imposed by subsampling were negligible, usually below 1%. For each dataset, an optimal subsampling point that improves prediction properties was observed, but the improvements were also negligible.

Conclusion: Combining subsampling with Gibbs sampling is an interesting ensemble algorithm. The investigation indicates that the subsampling bootstrap Markov chain algorithm substantially reduces computational burden associated with model fitting, and it may slightly enhance prediction properties.

Keywords: Genome-wide selection, Bayesian analysis, Bootstrapping

Background

The use of genomic tools has become important for the genetic improvement of complex traits in plants and animals through genome-wide prediction (GWP). GWP provides an interesting solution for the selection of traits with low heritability, such as grain yield in crops and milk production in dairy cattle, as well as for traits that present challenging or expensive phenotyping.

Over the past decade, researchers have tried to overcome the pitfalls of increased computational burden associated with gains in prediction accuracy from GWP of complex traits. Increases in predictive ability (and computational burden) are often associated with better statistical learning properties, such as regularization and variable selection [1]. Hence models with an improved ability to identify patterns provide more robust predictions, but computational costs are involved.

In statistical learning, resampling techniques are common approaches used to turn weak learners into strong learners [2]. Gianola et al. [3] showed that bootstrapping aggregation could improve prediction accuracy of kernel-based genomic best linear unbiased prediction (GBLUP) model in genomic prediction of plant and animals. We hypothesized that a similar approach could apply to whole-genome regression methods, often referred to as the Bayesian alphabet [4].

Besides computational advantages offered by some resampling methods, these techniques may also help to overcome theoretical shortcomings of some of these Bayesian methods, such as the bias of BayesA [5]. The objective of this study was to evaluate the predictive and computational outcomes from the application of a resampling technique ensemble with the Gibbs sampler to a Bayesian ridge regression model.

Sampling procedures

In addition to the increasing number of markers available over time due to higher density single nucleotide polymorphism (SNP) arrays and even resequencing, computation challenges include the large number of samples from which those genotypes are taken [6]. The computational burden associated with large population

* Correspondence: krainey@purdue.edu
[1]Department of Agronomy, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN 47907, USA
Full list of author information is available at the end of the article

Xavier *et al. BMC Bioinformatics* (2017) 18:191

Page 2 of 7

sizes is more evident in plant breeding, where hundreds of crosses with large offspring are genotyped and selected every season using GWP. Sampling methods are often necessary to enable such complex statistical procedures in large datasets. Among those, two main classes of sampling techniques are Markov chain Monte Carlo (MCMC) and Bootstrapping.

The MCMC method is possibly the most popular Monte Carlo algorithm with application to linear models, providing a feasible framework to resolve high-dimensional problems (i.e., more parameters than observations) with moderate computer power [7]. Likewise, bootstrapping also provides an interesting framework for solving large-scale problems [8, 9], particularly a method known as subsampling [10] used to reduce data dimensionality.

Gibbs sampling

Gibbs sampling is a widely used MCMC technique, applied in conjunction with Bayesian methods to generate the posterior distribution of the parameters. The posterior distribution is denoted as $p(\Theta|X)$, where Θ represents the set of unknown parameters $\Theta = \{\theta_1, , \theta_2, ..., \theta_r\}$, and X represents the data. The Gaussian model described in the following section, unknown parameters include the intercept (μ), the vector of regression coefficients ($\mathbf{b}$) and variance components, as $\Theta = \{\mu, \mathbf{b}, \sigma_b^2, \sigma_e^2\}$, whereas the observed data comprises the genotypic information ($\mathbf{X}$) of individuals and phenotype ($\mathbf{y}$), as $X = \{\mathbf{X}, \mathbf{y}\}$.

Gibbs sampling algorithms are based on updating each parameter with samples drawn from the full-conditional posterior distribution, one parameter at a time while holding every other parameter constant. Each parameter θ is sampled from

$$p(\theta|X) \propto f(X|\theta)\pi(\theta), \forall \theta \in \Theta, \quad (1)$$

where $p(\theta|X)$ denotes the posterior distribution of θ, the likelihood is expressed as $f(X|\theta)$ and the prior distribution of θ is $\pi(\theta)$.

In most implementations, regression coefficients are sampled individually from normal distributions whereas variance components are sampled from scaled inverse chi-squared distributions [4, 5]. Every time a parameter (i.e., regression coefficients and variance components) or a conjugated prior is updated, its value is stored as samples of the posterior distribution. The final Bayesian estimator is the expectation of the posterior distribution, obtained as the mean of the posterior distribution.

Bootstrapping aggregation

A natural strategy to increase prediction accuracy is to build and combine multiple prediction models generated from samples of a large dataset, averaging the outcome predictor [11]. Bootstrapping aggregation, or simply 'bagging', is implemented in linear models by fitting the function $f_1(x), f_2(x), ..., f_B(x)$ with B bootstrapped samples of the dataset and the final model, with reduced variance, will be given by

$$\hat{f}_{avg}(x) = \frac{1}{B} \sum_{b=1}^{B} \hat{f}_b(x), x \subset X. \quad (2)$$

Regression coefficients are stored each time the model is fitted, hence generating an empirical distribution of each parameter. Bagging parameters are obtained as the mean of this distribution.

With bootstrapping, when samples are obtained with replacement, the number of observations sampled is commonly the same size as the initial dataset, recognizing that some observations may be sampled more than once. When bootstrapping is performed with fewer samples than the original number of observations, sampling can proceed either with or without replacement. The latter case is known as subsampling.

Subsampling bootstrap Markov chains

MCMC and Bootstrap are usually implemented separately, such that some studies have attempted to compare the performance of these samplers [12]. Nevertheless, both methods can be co-implemented. A co-implementation that is becoming popular in the context of big data is a technique known as subsampling bootstrap Markov chain (SBMC). In this algorithm, the Markov chain update mechanism is performed upon a subset (x) of the whole data (X) and a different subset is used to update the parameters in each round of MCMC. Therefore, each parameter is sampled from the posterior distribution

$$p(\theta|x) \propto f(x|\theta)\pi(\theta), \forall \theta \in \Theta, x \subset X. \quad (3)$$

The concept of subsampling Gibbs sampler was first presented by Geyer [13] and some predictive properties were further investigated by MacEachern and Peruggia [14]. Regarding the applications to genome-wide prediction of complex traits, SBMC can be used to update the regression coefficients [15], hence increasing the computational performance of model fitting.

Methods
Statistical model
The family of whole-genome regression methods is a standard set of models widely applied for genomic prediction [4]. Among these, Bayesian ridge regression is a regularized model that assigns the same variance to every marker. The linear model is described as follows:

$$\mathbf{y} = 1\mu + \mathbf{Xb} + \mathbf{e} \quad (4)$$

where $\mathbf{y}$ is the response variable (i.e., the phenotypic information), μ is a scalar representing the intercept, $\mathbf{X}$ is

Xavier *et al. BMC Bioinformatics* (2017) 18:191

Page 3 of 7

the genotypic matrix coded as {0,1,2} for {AA, Aa, aa} where rows correspond to the genotypes and columns correspond to the molecular markers, $\mathbf{b}$ is a vector of regression coefficients that represents the additive value of allele substitutions, and $\mathbf{e}$ is the vector of residuals. In this model, both regression coefficients and residuals are assumed to be normally distributed as $\mathbf{b} \sim N(0, I\sigma_b^2)$ and $\mathbf{e} \sim N(0, I\sigma_e^2)$. The variances are assumed to follow a scaled inverse chi-squared distribution with a given prior shape (S) and prior degrees of freedom (ν), thus $\sigma_b^2 \sim \chi^{-2}(S_b, , \nu_b)$ and $\sigma_e^2 \sim \chi^{-2}(S_e, , \nu_e)$.

High-dimensional methods are regularized to enable fitting the model without losing predictive properties [2]. The regularization of linear models occurs by shrinking regression coefficients, which also biases predictions downwards [1]. The Bayesian ridge regression attempts to estimate regression coefficients with the minimum bias necessary for a satisfying prediction (i.e., minimum variance), a solution referred to as best linear unbiased predictor [4, 5]. As an optimization problem, the loss function to be minimized by the model (equation 4) that balances variance and bias is described as

$$L_2 = (\mathbf{y}-\mu-\mathbf{Xb})^{'}(\mathbf{y}-\mu-\mathbf{Xb}) + \lambda(\mathbf{b}^{'}\mathbf{b}) \tag{5}$$

where λ is the regularization parameter, the ratio between the residual variance and the genetic variance of marker effects, as $\lambda = \sigma_e^2/\sigma_b^2$. For the model in consideration, the regularization parameter assumes a single value for all regression coefficients.

Coefficient update

Sorensen and Gianola [16] show that the full conditional distribution of regression coefficients for Gibbs sampling from a normal distribution has a closed form. The expectation is derived from the Cholesky decomposition of the left-hand side (LHS) of the mixed model equation. The computational cost of operations such as solving the mixed model equation is described in terms of n observations and p parameters. The cost associated with the Cholesky decomposition is p^3, making it computationally unfeasible for high-dimensional problems ($p \gg n$), such as whole-genome regression methods. On the other hand, the Gauss-Seidel residual updating (GSRU) algorithm [15] has a computational cost of $3pn$, which is much lower than for the Cholesky decomposition in this scenario. A Gibbs sampler based on GSRU updates the jth regression coefficient as

$$b_{j^{t+1}}| * \sim N\left(\frac{x_j'e^t + x_j'x_j b_j^t}{x_j'x_j + \lambda_j}, \frac{\sigma_e^2}{x_j'x_j + \lambda}\right) \tag{6}$$

where x_j is the vector corresponding to the jth marker and $*$ represents the data and all parameters other

than the one being updated. The coefficient update is followed by update of the vector of residual

$$\mathbf{e}^{t+1} = \mathbf{e}^t + \mathbf{x}_j(b_j^{t+1}-b_j^t). \tag{7}$$

The greatest advantage of GSRU comes from the low computational cost of updating the right-hand side (RHS) of the mixed model equation [15], solving the linear system one parameter at a time without computing $\mathbf{X'X}$. Subsequently, variance components are updated as

$$\frac{\sigma_b^2| * \sim \mathbf{b}^{'}\mathbf{b} + S_b\nu_b}{\chi_{p+\nu_b}^2 \text{ and } \sigma_e^2| * \sim \mathbf{e}^{'}\mathbf{e}+S_e\nu_e} \atop \chi_{n+\nu_e}^2. \tag{8}$$

where S_e, ν_e, S_b, and ν_b correspond to the prior parameters "shape" and "degrees of freedom" of the residual and genetic variance, respectively.

SBMC extension

We here propose incorporating subsampling into the Gibbs sampler. This variation implies sampling a ψ fraction of the data ($\psi \in [0, 1]$) to update regression coefficients and residual variance in each round of MCMC.

For a matter of notation, let $\mathbf{X}^{\sim}$ and $\mathbf{e}^{\sim}$ represent the bagged subsamples, in other words, a fraction of $\mathbf{X}$ and $\mathbf{e}$ that contains ψ percent of observations sampled at random in a given round of MCMC. This modified GSRU would have an expected computational cost of $3pn\psi$.

To accommodate bagged samples, sampling algorithms of regression coefficients and residual variance undergo a slight modification. Regression coefficients are updated or sampled as

$$b_{j^{t+1}}| * \sim N\left(\frac{\mathbf{x}^{\sim'}_j \mathbf{e}^{\sim t} + \psi\mathbf{x}'_j\mathbf{x}_j b_j^t}{\psi\mathbf{x}'_j\mathbf{x}_j + \lambda_j}, \frac{\sigma_e^2}{\psi\mathbf{x}'_j\mathbf{x}_j} + \lambda_j\right) \tag{9}$$

with subsequent residual update

$$\mathbf{e}^{\sim t+1} = \mathbf{e}^{\sim t} + \mathbf{x}_j(b_{j^{t+1}}-b_j^t). \tag{10}$$

The entire k^{th} round of MCMC is updated using the subsampled dataset $x^k = \{\mathbf{X}^{\sim}, \mathbf{e}^{\sim}\}$. Since the residual variance is a function of the number of observations, its update is slightly modified from equation 8 as

$$\frac{\sigma_e^2| * \sim \mathbf{e}^{\sim'}\mathbf{e}^{\sim} + S_e\nu_e}{\chi_{\psi n+\nu_e}^2.} \tag{11}$$

The sampling procedure above assumes that the variance associated to markers in the subsamples are approximately the same as in the whole data ($\sigma_x^2 \approx \sigma_x^2$). That is, the marker sum of squares ($\mathbf{x'x}$) is expected to reduce linearly according to the proportion of bag

Xavier *et al. BMC Bioinformatics* (2017) 18:191

Page 4 of 7

samples (ψx'x) to avoid recalculating the sum of squares of bagged markers ($x^{\sim'} x^{\sim}$) for each round of MCMC. In genetic terms, the subset is assumed to have the same allele frequencies as the whole set.

The SBMC algorithm is implemented in the R package bWGR [17] using the R^2 rule proposed by Pérez and de Los Campos [18] to estimate prior shapes using the whole data, based on $R^2 = 0.5$, with the values of prior degrees of freedom set as $\nu_e = 5$ and $\nu_b = 5$. In the R^2 rule [18], prior shapes are estimated as

$$S_b = R^2 \times \sigma_y^2 \times (\nu_b + 2) / \sum_j \sigma_{x_j}^2 \qquad (12)$$

and

$$S_e = \left(1 - R^2\right) \times \sigma_y^2 \times (\nu_e + 2). \qquad (13)$$

Dataset
Three datasets available on R packages [18, 19] were chosen to demonstrate the effect of bagging on genomic prediction, including a wheat panel from the International Maize and Wheat Improvement Center (CIMMYT), as the median of grain yield observed in four environments [20]; a mouse panel designed to study body mass index [21] but using only half the SNP panel obtained by skipping every other marker; a soybean panel with eight bi-parental families with elite parents from the SoyNAM project [19] with phenotypes observed in eighteen environments; and a simulated F_2 population with 10 chromosomes of 50 cM each, genotyped at density of 1 SNP/cM, and with one QTL every 10 cM placed between markers. Heritability of traits was computed by restricted maximum likelihood (REML) upon the animal model with additive kernel [22]. Markers with minor allele frequency below 0.05 were removed. Datasets are summarized in Table 1.

Prediction metrics
Prediction statistics were obtained with a 10-fold cross validation scheme. We fitted the Bayesian ridge regression model using subsampling from 25 to 100%, by increment of 1%, with and without replacement. We set the algorithm for 4000 MCMC iterations to ensure convergence [16], with 500 of burn-in [18].

To determine the efficacy of subsampling, we evaluated the mean square prediction error (MSPE), prediction bias as the slope of linear regression between predictions and observations ($\beta_{y,y}$), computation time in minutes, and predictive ability as the Pearson's correlation between predictions and observations ($Cor_{y,y}$).

Results
The mean outcome of prediction metrics across datasets is presented in Fig. 1. The results by individual dataset are presented in the Additional file 1. Numeric results for some proportions of subsampling are presented in Table 2.

Computational improvement
The computational time had a linear response to subsampling (Fig. 1d). As expected, subsampling is clearly beneficial to speed up the computation of model fitting. The same trend was observed for individual datasets (Additional file 1). Although our evaluation of the improvement of computational performance used relatively small datasets, we believe the results must hold for larger datasets.

In comparison to fitting the model with whole data (Table 2), the computation time to fit the model at 50% subsampling was 33.6% faster with replacement and 58.3% faster without replacement. Yet, the computational cost was less than expected, once $3pn\psi$ with $\psi = 0.5$ should provide a model fitting 100% faster. This difference can be attributed to the computational cost of the sampling process along with the fixed cost of the initial problem settings. Computation time 100% faster was achieved for subsampling 33% (or less) without replacement.

Interestingly, subsampling with replacement presented a slightly higher computational cost, also presenting worse predictive properties for subsampling lower than 40% or higher than 60%.
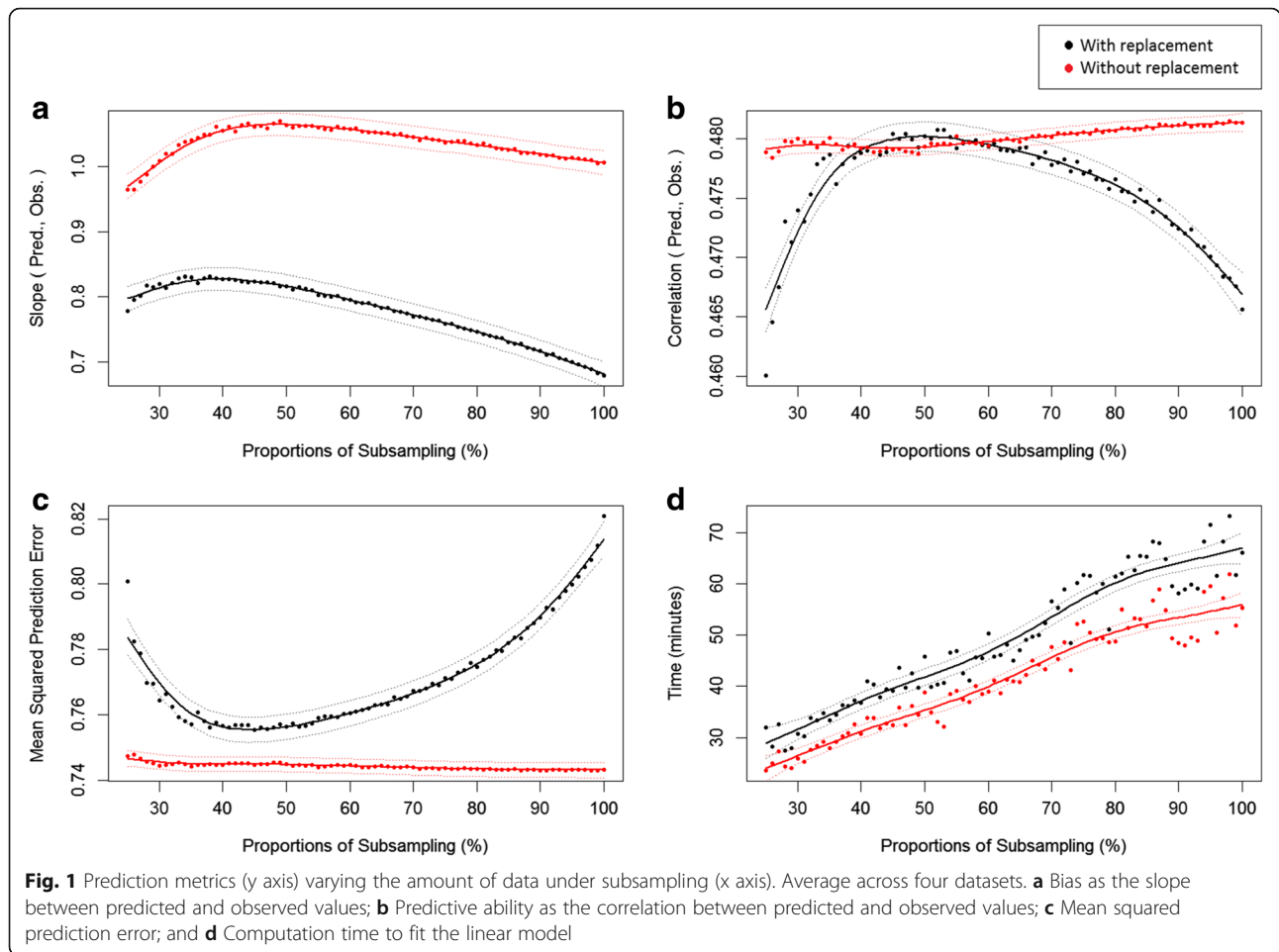
Implications of subsampling on prediction parameters
Bias
The use of the complete dataset was nearly unbiased (Table 2). Subsampling with replacement was biased downwards, presenting the least bias at 40% replacement ($\beta_{y,y} = 0.824$). Subsampling without replacement presented slight upward bias, being 1.8 and 5.8%

Table 1 Summary of datasets used in this study

Species	Population type	Trait	n	p	h^2	Source
Mouse	Heterogeneous stock	Body mass index	1814	5173	0.146	Legarra et al. [21]
Soybean	Nested Ass. Panel	Grain yield	1079	4307	0.345	Xavier et al. [19]
Wheat	Diverse panel	Grain yield	599	1209	0.434	Crossa et al. [20]
Simulation	Experimental F2	Simulated	400	500	0.516	Technow [23]

Fig. 1 Prediction metrics (y axis) varying the amount of data under subsampling (x axis). Average across four datasets. **a** Bias as the slope between predicted and observed values; **b** Predictive ability as the correlation between predicted and observed values; **c** Mean squared prediction error; and **d** Computation time to fit the linear model

more biased than the complete dataset at 33 and 50% subsampling, respectively.

Predictive ability

Across datasets (Table 2), the loss in predictive ability was negligible. Correlation between predictions and observations decreased 0.2% by subsampling with replacement at 50% subsampling, and 0.4% without replacement at both 33 and 50% subsampling.

Table 2 Summary of prediction metrics with for the complete dataset (Complete), and subsampling 50% with replacement (wR), and 33 and 50% without replacement (woR)

	Time (min.)	Cor$_{y,y}$	MSPE	$\beta_{y,y}$
Complete	55.90	0.4814	0.7431	1.0058
woR 33%	27.90	0.4794	0.7454	1.0239
woR 50%	35.32	0.4794	0.7447	1.0642
wR 50%	41.84	0.4802	0.7562	0.8161

Cor$_{y,y}$, correlation between observed and predicted value; MSPE, mean squared prediction error; $\beta_{y,y}$, Prediction bias

MSPE

The negative impact on MSPE due to subsampling was also negligible. An increase of 0.3 and 0.2% were observed at 33 and 50% subsampling without replacement (Table 2). The impact of subsampling on MSPE was slightly higher with replacement, increasing 1.76% at 50% subsampling.

Dataset specific analysis

Although negligible, we observed a slight improvement in predictive ability and MSPE for all datasets at some optimal subsampling point. The optimal subsampling and respective improvement in predictive ability and MSPE are presented in Table 3.

Discussion

Prediction machinery

Any algorithm that enhances prediction or computation performance is valuable for machine learning. At its optimal utilization, SBMC has the potential of improving prediction while reducing the computational cost [14]. However, reported results vary regarding any prediction

Xavier *et al. BMC Bioinformatics* (2017) 18:191

Page 6 of 7

Table 3 Optimal sampling observed for individual datasets to enhance predictive ability (PA) and mean squared prediction error (MSPE). Subsampling performed with (wR) and without replacement (woR)

	Optimal PA	Increase in PA	Optimal MSPE	Decrease in MSPE
Mouse	wR 66%	2.5%	woR 32%	<0.1%
Soybean	woR 25%	0.1%	woR 25%	0.1%
Wheat	woR 34%	0.7%	woR 33%	0.5%
Simulated F_2	wR 87%	0.1%	wR 66%	0.2%

improvement provided by subsampling [8, 24]. Subsampling has not been investigated in big data, for neither large n nor large p, and that is a specific niche where subsampling may work best.

Previous studies indicate that there are no guarantees that SBMC will improve prediction, but it at least provides results equivalent to the whole dataset; however, we showed that subsampling can also provide a positive outcome for genomic prediction besides the computational aspects (Table 3), where the improvement reached 2.5% for the mouse data. We recommend including a bagging WGR with 50% subsampling without replacement in cross-validation studies looking for the most accurate prediction model.

Random data

An interesting statistical property provided by SBMC is that data is sampled from a larger set, which is associated with that definition of a random term. This occurs because the observations used to update parameters are sampled from the empirical distribution of the data. This property violates the Bayesian assumption that data are *given*.

In classical Bayesian analysis, inferences are made based upon the posterior distribution of *parameters given data*, whereas random data implies that the parameters are sampled from the distribution of parameters given the current state of data. MCMC drives the posterior towards a relative entropy, possibly with larger sample variance associated with the continuous resampling used to update parameters with different subsets of data, but without obvious implications for the interpretation of the results [25].

Incompleteness of data

Geyer [26] discussed the issue of subsampling Markov chains concluding "one does not get a better answer by throwing away data." Nevertheless, he emphasizes that the value of the technique is 1) the reduction of dimensionality of n, and 2) the reduction of auto-correlation among chains.

Our counterargument is that the all data are used in the course of model fitting, although not simultaneously. In addition, accurate estimates are obtained when the subsampling strategy is used correctly [14]. We show

that subsampling is a valid approach for genomic prediction purposes to fit high-dimensional models ($p \gg n$).

Future directions

Subsampling uses only part of the data to fit the model in each MCMC round, that enables the computation of prediction statistics with the subset left out, which is referred to as out-of-bag statistics (OOB) [27]. The information provided by OOB is similar to the outcome of a cross-validation, with the advantage of being computed during the model fitting. Therefore, OOB could be used to re-weight observations (i.e., boosting). Another possibility is to adapt SBMC to other learning methods, such as elastic net [28], where OOB statistics could be utilized in the search for the tuning parameters without having to perform explicit cross-validation [29].

Conclusion

SBMC decreases computation time without compromising prediction properties. We observed that subsampling approximately 33–50% without replacement and 40–60% with replacement in each round of MCMC is advantageous for fitting the model. Subsampling can dramatically reduce computational burden with little reduction in accuracy and, in some cases, enhanced predictive properties. This study provides insight into a general method for incorporating a particular type of bagging ensemble into the Gibbs sampling of whole-genome regressions.

Additional file

Additional file 1: Results presented by individual dataset **Figure S1**. Time to fit the model (y axis) varying the subsampling method (x axis). **Figure S2**. Prediction ability (y axis) varying the subsampling method (x axis). Methods include Bayesian ridge regression (BRR) with regular sampler, and SBMC subsampling from 25 to 100%, with and without replacement. **Figure S3**. Mean squared prediction error (y axis) varying the subsampling method (x axis). Methods include Bayesian ridge regression (BRR) with regular sampler, and SBMC subsampling from 25 to 100%, with and without replacement. **Figure S4**. Bias (y axis) varying the subsampling method (x axis). Methods include Bayesian ridge regression (BRR) with regular sampler, and SBMC subsampling from 25 to 100%, with and without replacement. (DOCX 232 kb)

Abbreviations

CIMMYT: International Maize and Wheat Improvement Center; GBLUP: Genomic best linear unbiased prediction; GSRU: Gauss-Seidel residual updating; GWP: Genome-wide prediction; LHS: Left-hand side; MCMC: Markov chain Monte

Xavier *et al. BMC Bioinformatics* (2017) 18:191

Page 7 of 7

Authors' contribution
AX wrote the manuscript. SX revised the mathematical notation and theoretical basis of subsampling. WM and KMR provided insight for the method evaluation, predictive metrics and the suggested the datasets. 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.

Publisher's Note
Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Author details
[1]Department of Agronomy, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN 47907, USA. [2]Department of Plant Science, University of California, 3134 Batchelor Hall, Riverside, CA 92521, USA. [3]Department of Animal Science, Purdue University, 915 W. State St., Lilly Hall, West Lafayette, IN 47907, USA.

References
1. Okser S, Pahikkala T, Airola A, Salakoski T, Ripatti S, Aittokallio T. Regularized machine learning in the genetic prediction of complex traits. PLoS Genet. 2014;10(11):e1004754.
2. Hastie T, Tibshirani R, Friedman J. Elements of statistical learning. New York: Springer; 2009.
3. Gianola D, Weigel KA, Krämer N, Stella A, Schön CC. Enhancing genome-enabled prediction by bagging genomic BLUP. PLoS One. 2014;9(4):e91693.
4. de los Campos G, Hickey JM, Pong-Wong R, Daetwyler HD, Calus MP. Whole-genome regression and prediction methods applied to plant and animal breeding. Genetics. 2013;193(2):327–45.
5. Gianola D. Priors in whole-genome regression: the Bayesian alphabet returns. Genetics. 2013;194(3):573–96.
6. Misztal I. Inexpensive computation of the inverse of the genomic relationship matrix in populations with small effective population size. Genetics. 2016;202(2):401–9.
7. Scott SL, Blocker AW, Bonassi FV, Chipman HA, George EI, McCulloch RE. Bayes and big data: the consensus monte Carlo algorithm. Int J Manag Sci Eng Manag. 2016;11(2):78–88.
8. Flegal JM. Applicability of subsampling bootstrap methods in Markov chain monte Carlo. In: Monte Carlo and quasi-monte Carlo methods. Heidelberg: Springer; 2012. p. 363–72.
9. Kleiner A, Talwalkar A, Sarkar P, Jordan M. The big data bootstrap. arXiv preprint arXiv;1206.6415;2012.
10. Politis DN, Romano JP, Wolf M. On the asymptotic theory of subsampling. Statistica Sinica. 2001;11(4):1105–24.
11. James G, Witten D, Hastie T, Tibshirani R. An introduction to statistical learning. New York: Springer; 2013. p. 331.
12. Alfaro ME, Zoller S, Lutzoni F. Bayes or bootstrap? a simulation study comparing the performance of Bayesian Markov chain monte Carlo sampling and bootstrapping in assessing phylogenetic confidence. Mol Biol Evol. 2003;20(2):255–66.
13. Geyer CJ. Practical Markov chain monte Carlo. Stat Sci. 1992;7(4):473–83.
14. MacEachern SN, Peruggia M. Subsampling the gibbs sampler: variance reduction. Stat Probab Lett. 2000;47(1):91–8.
15. Legarra A, Misztal I. Technical note: Computing strategies in genome-wide selection. J Dairy Sci. 2008;91(1):360–6.
16. Sorensen D, Gianola D. Likelihood, Bayesian, and MCMC methods in quantitative genetics. Springer Science & Business Media, New York. 2002.
17. Xavier A, Muir W, Rainey KM. bWGR: Bagging Whole-Genome Regression. CRAN, version 1.3.1. 2016.
18. Pérez P, de Los Campos G. Genome-wide regression & prediction with the BGLR statistical package. Genetics. 2014;198(2):483–95.
19. Xavier A, Beavis WD, Specht JE, Diers BW, Howard R, Muir WM, Rainey KM. SoyNAM: Soybean Nested Association Mapping Dataset. CRAN, version 1.2. 2015.
20. Crossa J, de Los Campos G, Pérez P, Gianola D, Burgueño J, Araus JL, Makumbi D, Singh RP, Dreisigacker S, Yan J, Arief V. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics. 2010;186(2):713–24.
21. Legarra A, Robert-Granié C, Manfredi E, Elsen JM. Performance of genomic selection in mice. Genetics. 2008;180(1):611–8.
22. Xu S. Mapping quantitative trait loci by controlling polygenic background effects. Genetics. 2013;195(4):1209–22.
23. Technow F. hypred: Simulation of Genomic Data in Applied Genetics. CRAN, version 0.5. 2014.
24. Brooks S, Gelman A, Jones G, Meng XL. editors. Handbook of Markov Chain Monte Carlo. CRC Press. 2011.
25. Shalizi CR. Dynamics of Bayesian updating with dependent data and misspecified models. Electron J Stat. 2009;3:1039–74.
26. Geyer CJ. Introduction to Markov chain Monte Carlo. Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC. 2011;10:3–48.
27. Breiman L. Out-of-bag estimation. Technical Report, Statistics Department, University of California Berkeley, Berkeley CA 94708. 1996b;33,34; 1996.
28. Zou H, Hastie T. Regularization and variable selection via the elastic net. J R Stat Soc Series B Stat Methodology. 2005;67(2):301–20.
29. Xavier A, Muir WM, Craig B, Rainey KM. Walking through the statistical black boxes of plant breeding. Theor Appl Genet. 2016;129(10):1933–49.