IMPUTATION OF MISSING GENOTYPES IN HIGH DENSITY SNP DATA

G. Moser[1], M.S Khatkar[2] and H.W. Raadsma[2]

The CRC for Innovative Dairy Products
[1] Bellbowrie, QLD, 4070, Australia
[2] ReproGen, Faculty of Veterinary Science, The University of Sydney, Camden, Australia

SUMMARY

The accuracy and computational complexity of five methods to impute missing genotypes in high density SNP data was investigated. The haplotype reconstruction package fastPHASE reached the highest accuracies (91% to 98%) for varying proportions (0.2% to 8%) of missing genotypes. Alternative methods based on principal component analysis were less accurate (67% to 94%), but their computational demand was an order of magnitude lower.

INTRODUCTION

Missing genotype information is a common feature of high density SNP datasets. Even if the missing rate is low, eliminating SNP markers with incomplete observations will result in considerable loss of information. Several methods exist to deal with missing genotypes, such as imputing missing values of row averages or medians, but accuracy can be significantly improved by exploiting the correlation between data. Methods using haplotype reconstruction informed by linkage disequilibrium between SNP are commonly used to infer missing genotypes. However, when the number of loci is large, these approaches are computationally demanding. Less accurate but faster methods exist, that use the global covariance structure of SNP markers on a chromosome. The purpose of this study was to compare the performance of haplotype-based with covariance-based imputation methods.

MATERIALS AND METHODS

Missing genotype imputation. Five methods were used to infer missing values (MV) in SNP data. The haplotype reconstruction package fastPHASE (Scheet and Stephens 2006) uses a Hidden Markov Model to describe the spatial distribution of clusters of haplotypes along the chromosomes. FastPHASE uses the EM algorithm to estimate genetic parameters and haplotype frequencies from which missing genotypes are inferred. FastPHASE requires the specification of 5 tuning parameters to control the algorithm, which could be inferred through cross-validation. We chose two values for the number of haplotype clusters (K parameter) viz. 10 (fastPHASE1) and 50 (fastPHASE2), the default values were used for the other parameters. For principal component analysis (PCA) based imputation, we employed four algorithms implemented in the freely available R package pcaMethods (R Development Core Team 2007, Stacklies *et al.* 2007); Probabilistic PCA (PPCA), Bayesian PCA (BPCA), svdImpute (SVDI) and Nipals PCA (NIPALS). The principle behind the first three approaches is that missing values are initially set to the row averages, and singular value decomposition of the SNP matrix is used to create orthogonal principal components. The principal components, which correspond to the largest eigenvalues are then used to reconstruct the missing SNP genotypes in the SNP matrix. Nipals PCA uses the NIPALS algorithm (non-linear estimation by iterative partial least squares, Wold *et al.* 1966) for finding the principal components of the SNP matrix. In a study using gene expression data, Brock *et al*. (2008) found that covariance-based imputation methods are highly competitive with each other, but that no method was uniformly superior across different data sets and that the optimal method depended on the correlation structure of the data. The optimal number of principal

components is the only tuning parameter required by the methods. We fitted models varying the number of principal components in steps of 20. The best model was chosen as the one that provided the smallest prediction error.

Assessment of performance. The normalised mean root square error of prediction (NRMSEP) was used as the metric to compare the performance between methods. The NRMSEP normalises the square difference between observed and estimated values for a certain SNP locus by the variance within this locus. For the PCA methods, estimates are on a continuous scale and can be less than 0 or larger than 2. An ad-hoc binning algorithm was used to assign the inferred values into distinct genotypes, where genotype 0 was assigned if the estimate was ≤ 0.5, 2 if the estimate was > 1.5, and 1 when the estimate was > 0.5 and ≤ 1.5. Accuracy was computed as the percentage of correctly imputed genotypes, the misclassification rate was calculated as the proportion of incorrectly inferred MV on the total number of genotypes in the sample.

Data. Two SNP datasets for bovine chromosome 6 (BTA6) generated by the CRC for Innovative Dairy Products were used. The original data contained 0.8% and 0.6% missing values. The first set (BTA6.1) consisted of a sample of 377 bulls genotyped for 1446 SNPs. Missing values were generated for a random selection of 25% of all loci. We randomly masked 1%, 5%, 10% and 25 % of the genotypes at the chosen loci. This corresponds to proportions of 0.25%, 1.52%, 3.14% and 7.97% newly generated MVs on the total number of genotypes. The second dataset (BTA6.2) comprised 1943 bulls genotyped for 325 SNPs. One locus was randomly selected in each segment of 50 consecutive SNPs and 10% of genotypes were assigned missing values at random, so that a total of 0.21% of all genotypes were generated as missing.

RESULTS AND DISCUSSION

Accuracy. Table 1 shows the summary statistic of the performance of the methods when we imputed an increasing proportion of genotypes for set BTA6.1. For all methods the prediction error increases with an increase in the proportion of missing genotypes. FastPHASE with tuning parameter setting 2 (fastPHASE2) was the most accurate method, with a prediction error about half the size compared to PCA methods. The number of correctly inferred genotypes was very high ranging from 98.3 % (MV=1385) to 97.3% (MV=43435). Of the PCA methods, NIPALS performed best for lower proportions of MVs, but PPCA and SVDI were more accurate if 8% of SNP genotypes were imputed. The choice of tuning parameters had a large impact on the accuracy of fastPHASE. Using fastPHASE with less optimal tuning parameters decreased the accuracy by about 6%. The number of misclassified SNP genotypes increased with the proportion of imputed genotypes. For low rates of missing genotypes (<3%) the percent of misclassified SNPs on the total number on genotypes was less than 0.5%, with the exception of BPCA. When 8% of genotypes were missing, fastPHASE2 had the lowest misclassification rate of 0.22%, which was 1% lower than the best PCA method (PPCA).

The methods were further compared for BTA6 in a second dataset (BTA6.2, Table 1) comprising a larger number of bulls (N=1943), but with a lower SNP density (N=325) in the chromosome. A total of 1333 missing values were generated, corresponding to a proportion of 0.21% of all genotypes. Again, fastPHASE2 was the most accurate method and assigned 95.8% of genotypes correctly. The best PCA methods were only slightly less accurate with NIPALS and PPCA imputing 94.05% and 93.15 of SNPs correctly. The chromosome-wide misclassification rate for NIPALS was less than 0.004% higher compared to fastPHASE2. The differences in accuracy between datasets may be an effect of the structure of the samples used and the density of

Table 1. Summary of performance of methods for missing genotype imputation

Data[1]	Missing SNP[2]	PPCA	BPCA	SVDI	NIPALS	fastPHASE1	fastPHASE2
				Prediction error (NRMSEP)			
	1385	0.546	0.756	0.555	0.501	0.513	0.206
	8304	0.571	0.817	0.581	0.536	0.527	0.288
	17103	0.579	0.833	0.588	0.554	0.523	0.288
	43435	0.606	0.887	0.618	0.636	0.531	0.286
				Accuracy (%)			
	1385	88.9	75.3	88.6	90.2	92.3	98.3
	8304	86.8	72.8	86.2	88.5	91.3	97.4
	17103	86.3	71.1	85.8	87.5	91.2	97.3
BTA6.1	43435	84.5	67.2	83.8	82.5	91.0	97.3
				Chromosome-wide misclassification rate(%)			
	1385	0.028	0.063	0.029	0.025	0.020	0.004
	8304	0.202	0.414	0.210	0.175	0.133	0.040
	17103	0.429	0.907	0.447	0.396	0.275	0.085
	43435	1.236	2.642	1.291	1.398	0.719	0.219
				Computing time			
	1385	0.58min	5.16min	4.25min	6.42min	1.15h	1.03d
	8304	1.21min	5.29min	4.19min	7.49min	1.18h	1.04d
	17103	2.59min	6.13min	4.18min	6.10min	1.00h	1.01d
	43435	6.42min	6.93min	4.21min	7.67min	1.06h	1.01d
				Prediction error (NRMSEP)			
	1333	0.145	0.190	0.156	0.145	0.162	0.123
				Accuracy (%)			
		93.2	91.7	86.7	94.1	91.5	95.8
BTA6.2				Chromosome wide misclassification rate (%)			
		0.015	0.028	0.018	0.013	0.018	0.008
				Computing time			
		7.09min	1.23min	5.08min	9.52min	1.11h	1.15d

[1] BTA6.1: 377 bulls x 1447 SNPs; BTA6.2: 1945 bulls x 325 SNPs. [2] Number of missing SNPs of 1436, 8304, 17102, 43435 and 1333 correspond to proportions of 0.25%, 1.52%, 3.14%, 7.97% and 0.21% missing values on the total number of genotypes.

SNPs on the chromosome. The data suggest that the gain in accuracy of haplotype-reconstruction methods is small for lower SNP densities. A reduction in gain in accuracy between fastPHASE and PCA methods was also found for shorter chromosomes (data not shown).

Computational complexity. The computational demand of the PCA methods is an order of magnitude lower than fastPHASE. For example imputation using fastPHASE with a tuning value of 50 haplotype clusters took more than 1 day to complete, whereas the all PCA based methods required less than 10min computing time. The estimates of PPCA, BPCA and SVDI are based on the routines implemented in the R package pcaMethods. Their speed depends largely on the computation of the singular value decomposition of the SNP matrix and implementation of these algorithms using faster programming languages will improve speed substantially. The NIPALS routine was implemented in FORTRAN and computation time was reduced by a factor of 25 compared to the pcaMethod implementation.

CONCLUSIONS

So far our analysis shows that haplotype reconstruction methods like fastPHASE, and presumably similar programs, provide the highest accuracy for inferring missing genotype data when compared to principal component based methods. The increase in accuracy might not be sufficiently large to justify its computational demand if the only purpose is to infer missing SNP genotypes. The use of fastPHASE might become prohibitive for bigger datasets, since the number of haplotype clusters normally increases with more animals in the sample and computation time increases quadratically with an increase in clusters. The accuracies of the methods depend on the choice of tuning parameters, and optimal values are usually found by cross-validation. Cross-validation adds considerably to the running time of fastPHASE. For PCA based methods the optimal number of principal components is the only tuning parameter required. The study should be extended to complete genomes and investigate the performance in relation to the accuracy of methods used to predict genomic breeding values.

ACKNOWLEDGMENTS

The genotype data was provided by the Co-operative Research Centre for Innovative Dairy products.

REFERENCES

Brock, G.N., Shaffer, J.R., Blakesley, R.E., Lotz, M.J. and Tseng, G.C. (2008) *BMC Bioinformatics* **9:**12.

R Development Core Team (2007) *Statistical Computing*, http://www.R-project.org.

Scheet, P. and Stephens, M. (2006) *Am J Hum Genet* **78**:629.

Stacklies, W., Redestig, H., Scholz, M., Walther, D. and Selbig, J. (2007) *Bioinformatics* **23**:1164.

Wold, H. (1966) In "Multivariate Analysis", p. 391, editor P.R. Krishnaiah, P.R., Academic Press, NY