Genotype Imputation Based on Discriminant and Cluster Analysis

Medhat Mahmoud, Theo Meuwissen and Thore Egeland
Norwegian University of Life Sciences, Ås, Norway
Correspondence to Medhat Mahmoud, E. mail: mahmoud@fbn-dummerstorf.de

Abstract

The recent development of high-throughput systems for genotyping SNP in Eukaryote has led to an extraordinary amount of research activity, particularly in areas such as whole-genome selection of livestock and genome-wide association studies for detection of quantitative trait. Recent technological advances allow us to rapidly genotype more than 10 million SNPs in an individual, accounting for 10% of the estimated number of common SNPs (more than 1% minor allele frequency) across the population. As a result of missing SNPs, true associations might be missed if the causal SNP is not genotyped or if the causal variant is an unknown variant. SNP imputation is important in reducing the cost of re-sequencing and when genotyping all considered animals may be too costly and sometimes not feasible because DNA may not be available for all animals. Computational algorithms and statistical methods have been developed to account for some of the unobserved variants. The main idea behind these methods is based on the observation that SNPs in close proximity to one another in the genome tend to be correlated, or in non-random association (linkage disequilibrium). "Several articles have described comparisons of imputation methods with respect to computational efficiency and the accuracy of results". Consequently, we perceived a substantial need to proposing a new technique for SNP Imputation with applying linear Discrimination and Clustering Analysis Algorithms. To evaluate the factors potentially affecting imputation accuracy rates (ARs), we used simulated data sets to investigate the effects of Linkage disequilibrium (LD), Minor allele frequency (MAF) of un-typed SNPs, marker density (MD), reference sample size (n) and the different numbers of SNPs in every haplotype block, in imputation accuracy rate (AR) and the performance of linear discriminant analysis and clustering Analysis as a SNP imputation method. In optimal state of genotype data (in High LD, low MAF, and high density haplotype blokes) both methods (Clustering and discrimination) were working efficiently, and the accuracy can reached 89 %.

Key words: SNP Imputation, Clustering, Linear discrimination

Introduction

Imputation is the substitution of some value for missing data, the practice of 'filling in' missing data with plausible values, is an attractive approach to analysing incomplete data. When substituting for a single value, it is known as "unit imputation"; when substituting for a component or a complete variable or item, it is known as "item imputation". After imputing all missing values, the dataset can then be technically analysed using normal methods for complete data. We should ideally take into our account that there is a greater degree of uncertainty than if the imputed values had actually been observed.

There are many reasons behind why the data is missing, one nature of missing data could be 'missing completely at random' (MCAR), and it may be because the equipment malfunctioned, or the data were entered in an uncorrected way. When some data are missing completely at random, it means that the probability that an observation X_i is missing is unrelated to the value of X_i or to the value of any other variables, e.g. Human HapMap would not be considered as MCAR if Whites were more likely to omit reporting genotype than African Americans. MCAR is an important consideration, because in this case the analysis remains unbiased. We may lose power for our design, but the estimated parameters are not biased by the absence of data. If data are not completely missing at random then they are classified as 'Missing Not at Random' (MNAR). When the data are MNAR then we have the problem of a biased

dataset, and the only way to obtain an unbiased estimate of parameters is to model the missing-ness or to write a model that accounts for the missing data (Dunning and Freedman 2008).

Genome-wide imputation

Recent technology in high-throughput genotyping estimated that the human genome contains more than 7 million common single nucleotide polymorphisms (SNPs) with minor allele frequencies (MAF) about 5% (Barrett JC., Cardon L. R., 2006), and only a small fraction of these SNPs can be directly assayed using current high-density microarrays. Due to the linkage disequilibrium (LD) among neighbours markers, many un-typed or missing SNPs are highly correlated with one or more surrounding nearby assayed SNPs. Therefore, testing assayed SNPs for association to traits of interest will have some power to detect or prediction of un-typed causal SNPs. Further, if the assayed SNPs are uniformly distributed across the genome, maximal genetic coverage can be achieved (Hao K., Schadt E. E., 2008). The same in genome-wide association studies, where significant signals suggest association between phenotypes and causal SNPs in the surveyed genome region. To improve this type of association analysis, the genotypes of missing SNPs can be imputed or predicted based on nearby markers (SNP) and then directly tested for association with phenotypes of interest (Servin B., Stephens M., 2007). The general aim of this study is to test the performance of modern multivariate techniques like (linear discriminant and clustering analysis) in SNP imputation. But in genotype data there are many factors that affecting the imputation accuracy, this will be investigated by

1- Testing linear discriminant imputation and clustering imputation in low and high Linkage disequilibrium genome regions (LD). 2- Testing linear discriminant imputation and clustering imputation in different levels of Minor allele frequency genome regions (MAF). 3- Testing linear discriminant imputation and clustering imputation in different levels Marker density regions (HD, LD). 4- Testing linear discriminant imputation and clustering imputation with different Reference sample sizes (n). 5- Testing linear discriminant imputation and clustering imputation with different Haplotype block sizes (K). N.B. We measure the Haplotype block size by counting the number of SNPs per haplotype block, not by Centimorgan.

Materials and Methods

Many datasets have been simulated for this study (See Table1), each Dataset consisted of a number of haplotype blocks (rows of individuals) and a number of SNPs markers (columns of variables), simulated with some constants parameters and only one varied parameter (parameter under investigation), for example: to investigate the effects of Minor allele frequency (MAF) of un-typed SNPs in imputation accuracy rate by a given imputation method, we simulate a different datasets with a constant correlation between SNPs, a constant reference sample size (n) and a constant number of SNPs in every haplotype blocks, but with a different levels of Minor allele frequency (MAF) of un-typed loci in each datasets, then we measured the differences in accuracy rate coming from using different dataset with different MAF.

Table 1. Presentation of all datasets used in imputation experiment.

Dataset	Test	Correlation	MAF %	No. haplotypes	No. SNP
1	No. of SNPs in **(LLD)** region	0.2	49	1000	Vary
2	No. of SNPs in **(HLD)** region	0.8	49	1000	Vary
3	Minor allele frequency **(MAF)**	0.2	Vary	1000	10
4	Marker density **(MD)**	Vary	49	1000	10
5	Reference sample size **(n)**	0.2	10	1000	10

Discriminant imputation

Let y be a categorical imputation variable with categories 1 and 2 and (x1, x2,…, xp) the set of predictor variables resulting from replacing any categorical predictor variable of y (major and minor allele) by its corresponding dummy variables (1 and 2). Let nj be the number of values of Yobs in category j, f $(.|\mu, \Sigma)$ the probability density function of the multivariate normal distribution with mean vector μ and variance Σ, respectively. Under the assumption that the conditional probability distribution of $x = (x_1, x_1, \dots, x_p)$ given y = j is a multivariate normal distribution with mean vector μj and covariance matrix Σ the underlying statistical model of discriminant imputation is given by

$$P(y = j|x) = \frac{f(x|\mu_j; \Sigma_j)\pi_j}{\sum_{v=0}^{s-1} f(x|\mu_v; \Sigma_v)\pi_v}$$

The previous model follows directly from substitution of $P(x|y = v) = f(x|\mu_v; \Sigma_v)$ and $P(y = v) = \pi_v$ into the formula of Bayes.

Nearest-neighbour (Clustering-based Imputation)

Nearest-neighbour imputation method (NIM) is an alternative form of hot-deck donor imputation. With this imputation, values from one record (the "donor") are used to replace the erroneous and missing values in another record (the "recipient"). The name "hot-deck" indicates that the donor and the recipient come from the same data set. Only records that are error-free may be used as donors.

To apply nearest-neighbour hot-deck imputation, a distance function D(i,k) must be defined that the measures the distance between two units i and k, where i is the item non-respondent and k is an arbitrary item respondent. The distance function D(i,k) can be defined in many different ways. A frequently used general distance function is the so called Minkowski distance:

$$D(i, k) = \left(\sum_j |x_{ij} - x_{kj}|^z\right)^{\frac{1}{z}}$$

where the x variables are numerical, and the sum is taken over all auxiliary variables; xij (xkj) denotes the value of variables xj in record i (k). Let the smallest value of D(i,k) be attained for item respondent d [d=arg mink D(i,k)], then respondent d is said to be the nearest-neighbour of the item non-respondent i and becomes its donor. For z=2 the Minkowiski distance is the Euclidean distance and for z=1 it is the so-called city-block distance. For larger z, large difference between xij and xkj are "punished" more heavily. In this Study we will use the Euclidean distance.

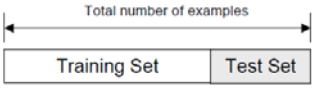
Practically, we divided the dataset (including the records with missing values) into (n) clusters. Next, missing values of an instance i are patched up with the plausible values generated from K's cluster. The following experiments will test the performance of the proposed method in genotype imputation task.

Validation (The holdout method)

The holdout method is the simplest method of cross validation. Each data set is split into two parts or sets, called the training-dataset (reference data-set) and the test-dataset. In LDA the prediction model is fit using the training data-set only. But in Clustering the prediction model is fit using the training data-set and the Test data-set. Usually we using 50% training data-set in this study, except in the last experiment where we measuring the effect of the size of the training dataset (where we varying the size of the training dataset (n)).

Then the same model is used to predict the outcome values for the data in the test data-set (only in LDA, were it has never seen these output values before).

The errors it makes (when we using the model to predict the outcome) are accumulated to give the mean absolute test set error, which is used to evaluate the model, in other word, the accuracy of this model counted by measuring the correlation between the true and the predicted value of the imputed SNP vector.



The advantage of this validation method is that It gives us the possibility to measure how much the size of the training-dataset (reference data-set) can affect the imputation accuracy, because in real life usually the data set contain some completed data (which can considered as training-dataset) and the rest have some missing values (considered as test-dataset)

However, the evaluation usually depends heavily on which data points end up in the training-dataset and which end up in the test-dataset. And **estimating the error rate** will be misleading if we happen to get an "unfortunate" split.

Results

Comparison between the performance of LDA and Clustering analysis in SNP imputation.

1- **Figure 1:** shows the effects of size of haplotype block (number of SNPs per haplotype), on imputation accuracy rate (AR) using low and high linkage disequilibrium dataset (LLD, HLLD). When LDA is used for imputation with constant MAF =49% and low linkage disequilibrium data the accuracy rate ranging from 60% (using 5 SNPs) to 70% (using 100 SNPs), while with High linkage disequilibrium data the accuracy rate ranging from 88% (using 5 SNPs) to 93% (using 100 SNPs). This is a high LD dataset AR is generally substantially higher and there is less improvement by increasing the number of SNPs.

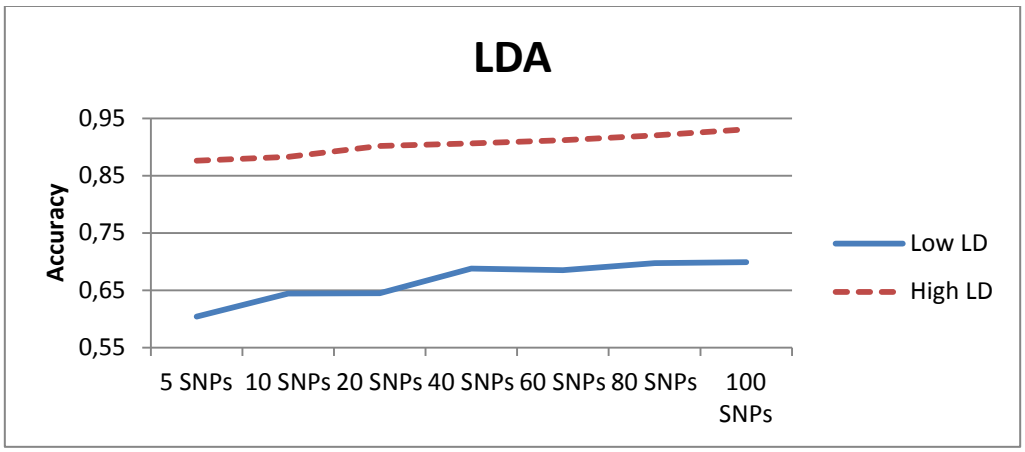


Figure 1. The effects of using Low and High linkage disequilibrium dataset on Accuracy rate of LDA in imputation.

2- **Figure 2:** Shows the effects of number of SNPs surrounding the missing one, in imputation accuracy rate (AR) using low and high linkage disequilibrium dataset (LLD, HLD). When clustering is used for imputation with constant MAF =49% and low linkage disequilibrium data the accuracy rate ranging from 55% (using 5 SNPs) to 71% (using 100 SNPs), while with High linkage disequilibrium data the accuracy rate ranging from 75% (using 5 SNPs) to 91% (using 100 SNPs). Generally clustering is less accurate than LDA and need more SNPs to reach high accuracy.

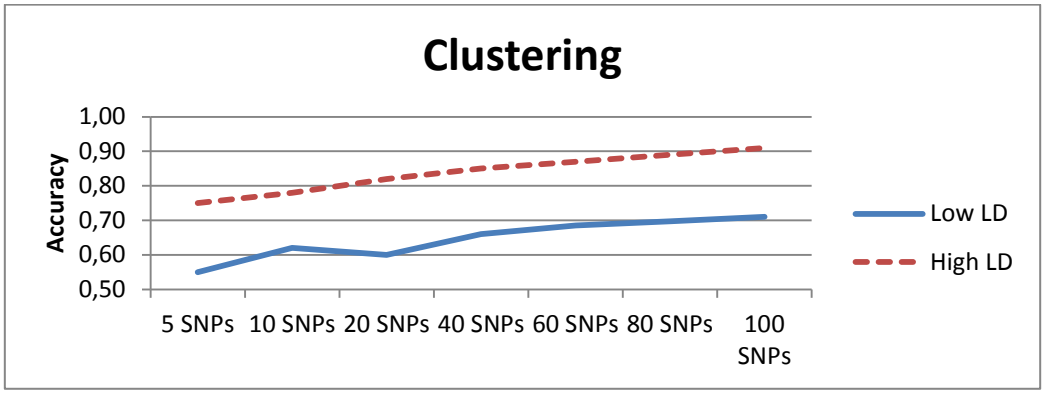


Figure 2. The effects of no.SNPs using Low and High linkage disequilibrium dataset on accuracy rate of Clustering in imputation.

3- **The effects of Minor allele frequency (MAF):** Figure 3.
Using LDA with constant correlation between SNPs = 0.10 and using the surrounding 10 SNPs the accuracy rate ranging from 0.99 (using MAF=0.10) to 0.75(using MAF=0.49), (**See Figure 3**). While using Clustering with constant correlation between SNPs = 0.10 and using the surrounding 10 SNPs the accuracy rate ranging from 92% (using MAF=0.10) to 69% (using MAF=0.49). It seems that AR is much more accurate when MAF is low compared to when it is high. A lower MAF usually corresponds to a stronger LD with nearby markers and the recombination plays a primary role in LD decay (Yu-Fang Pei., 2008).

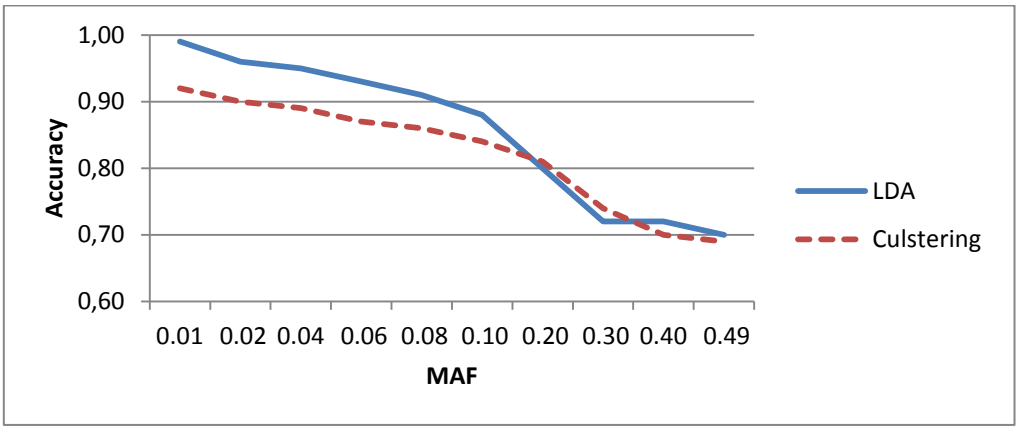


Figure 3. The effects of Minor allele frequency on accuracy rate using LDA and Clustering.

4- **The effects of marker density (MD):** Figure 4.
Using LDA With constant MAF =50% and using the surrounding 10 SNPs the accuracy rate ranging from 61% (using Corr. =0.10) to 97% (using Corr. =0.90). (**See Figure 4**) While using Clustering with constant MAF =49% and using the surrounding 10 SNPs the accuracy rate ranging from 55% (using Corr. =0.10) to 94% (using Corr. =0.90). Here, we measure the effect of Marker density by varying the correlation between markers (SNPs).

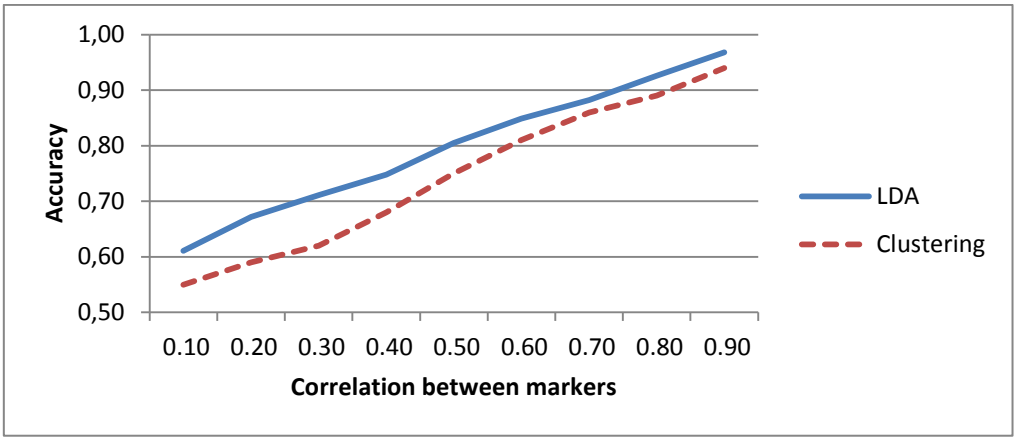


Figure 4. The effects of marker density on accuracy rate using LDA and Clustering.

5- **The effects of reference sample size (n):** Figure 5.
Using LDA, with constant MAF =10%, constant correlation between SNPs = 0.10 and using the surrounding 10 SNPs the accuracy rate ranged from 72% (using n =0.10) to 83% (using n =0.90), while using Clustering, with constant MAF =10%, constant correlation between SNPs = 0.10 and using the surrounding 10 SNPs the accuracy rate ranged from 33% (using n =0.10) to 79% (using n =0.90). This shows that clustering needs higher (n) to reach high accuracy.

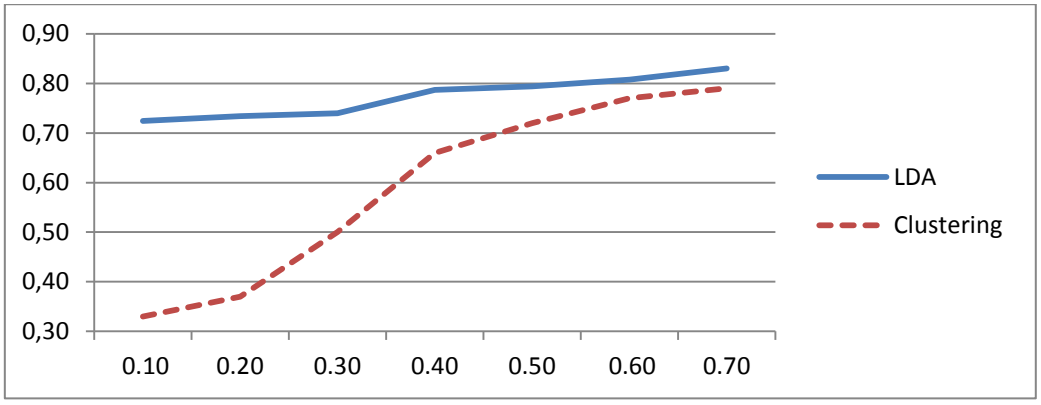


Figure 5. The effects of reference sample size on accuracy rate using LDA and Clustering.

5. Discussion and Conclusion

This study compared two different approaches (Discriminant-based SNP imputation and Nearest-neighbour or Clustering-based SNP imputation) using haplotype blocks instead of individual markers or all available markers. The average number of SNPs per haplotype blocks varying from 5 SNPs (in low LD region) to 100 (in High LD region). To investigate the performance of these two methods we simulated a group datasets each one simulated to test the effects of Linkage disequilibrium (LD), Minor allele frequency (MAF) of un-typed SNPs, marker density (MD), reference sample size (n) and the different numbers of SNPs in every haplotype block, in imputation accuracy rate (AR) and the performance of The Linear discriminant analysis and Clustering Analysis as a SNP imputation method. The dataset was also split in a training dataset and test dataset. The methods were validated using the holdout method then measuring the correlation between the true and imputed SNP in test dataset.

The performance of the elementary imputation methods, clustering and discrimination is generally good. However, to compare the performance of each algorithm with the currently used methods like in MACH, BEAGLE, and IMPUTE, more test experiments are needed to be conducted. Furthermore, to be sure that the algorithms are reliable, the same data sets should be used to run the experiments. Like any simulation study, this one has its limitations and advantages in some cases like:

1- In low LD region, the clustering-based method can use the correlation between records instead of the correlation between markers in the imputation process.
2- The Discriminant-based method also can handle numerical and categorical data simultaneously without rounding-up the results (which can affect the accuracy of imputation).

But in optimal state of genotype data (in High LD, low MAF, and high density haplotype blokes) both methods (Clustering and discrimination) were working efficiently, and the accuracy can reached 89 %.

Further studies and experiments are necessary before one can conclude whether the establishment of Discriminant-based and Clustering-based SNP imputation is feasible or not.

The Clustering-based SNP imputation models show a lot of promise for SNP imputation (and in Microarray analysis in general) based on the associations between records instead of using the association between markers.

Results obtained had many similarities with those obtained both from Discriminant-based imputation and Clustering-based SNP imputation approaches in similar datasets.

Linear discrimination can be considered as a complement algorithm for Clustering especially when applied to noisy data in what we can call "Cluster-based pattern discrimination CPD", which differs from standard clustering by being simultaneous subspace selection via linear discriminant analysis (LDA).

LDA is the most widely used in the two dimensions or categorised data. However, both statistical methods suffer from some deficiencies. Clustering analysis has the problem of selecting different values of K (i.e. number of nearest neighbouring haplotype records). Using different K-values results in different performance of the algorithms which in turn affects the final evaluation for the method accuracy. So that we propose to test the optimal K-value each time the algorithm is used. Finally, searching for a new technique and a new application or a new demonstration of Discriminant and Clustering analysis was the main interest of this thesis because nowadays the application of the modern statistical techniques such LDA, Clustering, PCA, PLS …and etc., are so important considerations in the field of Bioinformatics and Applied statistic.

References

Abramowitz, M. & stegun, I.A. 1964. *Handbook of Mathematical function*. U.S. Department of Commerce, national Bureau of Standard Mathematical Series, *55*.

Adriaans, P. & Zantinge, D. 1996. *Data Mining.* Harlow, England: Addison-Wesely.

Price, A.L., Patterson, N.J., Plenge, R.M. & Reich, D. 2006. *Principal components analysis corrects for stratification in genome-wide association studies.* Nature Publishing Group.

Bryk, A.S. & Raudenbush, S.W. 1992. *Hierarchical Linear Models*. Sage, Newbury Park.

Celeux, G. & Govaert, G. 1995. "Gaussian Parsimonious Clustering Models" *Pattern recognition 28*, 781-793.

Cormack, R.M. 1971. "A Review of Classification (with discussio)." *Journal of the Royal Statistical Society (A) 134:3,* 321-367.

Everitt, B.S., Landau, S. & Leese, M. 2001. *Cluster Anlysis (4th ed).*London: Hodder Arnold.

Fraley, C. & Raftery, A.E. 2002. "Model-Based Clustering, Discriminant Anlysis and Density Estimation." *Journal of the American statistical Association 97,* 611-631. *Statistics 49,* 297-310 (2000).

Greenacre, M.J. 1984. *Theory and applications of Correspondence Analysis.* London: Academic Press.

Hand, D., Mannila, H. & Smyth, P. 2001. *Principles of Data Mining*. Cambridge, MA: Mit Press.

Meng, X.L. 1995. Multiple-imputation inferences with uncongenial sources of input (with discussion). *Statistical Science 10,* 538-573.

Morgan, B.J.T. & Ray, A.P.G. 1995. "Non-uniqueness and inversions in Cluster Analysis." *Applied Statistics 44:1,* 117-134.

Solberg, T.R., Sonesson, A.K., Woolliams, J.A. & Meuwissen, T.H.E. 2009. *Reducing dimensionality for prediction of genome-wide breeding values:* Norwegian University of Life Sciences.

Pyle, D. 1999. *Data Preparation for Data Mining*. San Francisco: Morgan Kaufmann.

Shepard, R.N. 1980. "Multidimensional Scaling, Tree-Fitting, and Clustering." *Science 210:4468,* 390-398.

Sibson, R. 1978. "Studies in the Robustness of Multidimensional Scaling" *Journal of the Royal Statistical Society (B) 40,* 234-238.

Johnson, R.A. & Wichern, D.W. 2007. *Applied Multivariate Statistical Analysis* (Sixth Edition). Person Education, Inc.

Rubin, D.B. 1976. Inference and missing data. *Biometrika 63,* 581-592.

Rubin, D.B. 1987. *Multiple Imputation for Nonresponse in Surveys*. J. Wiley & Sons, New York.

Rubin, D.B. 1996. Multiple imputation (with discussion). *Journal of the American Statistical Association 91,* 473-489.

Schafer, J.L. 1997. *Analysis of Incomplete Multivariate Data*. Chapman & Hall, London.

Schafer, J.L. 1999. Multiple imputation: a primer. *Statistical Methods in Medical Research*, in press.

Takane, Y., Young, F.W. & De Leeuw, J. 1977. "Non-metric Individual Differences Multidimensional Scaling: Alternating Least Squares with optimal Scaling Features." *Psycometrika 42,* 7-67.

Ward, Jr., J.H. 1963. "Hierarchical Grouping to Optimize an Objective Function." *Journal of the American Statistical Association 58,* 236-244.

Westphal, C. & Blaxton, T. 1998. *Data Mining Solutions: Methods and Tools for Solving Real World Problems* (Paperback). New York: John Wiley.

Whitten, I.H. & Frank, E. 2005. *Data Mining: Practical Machine Learning Tool and* Techniques (2nd ed.) (Paperback). San Francisco: Morgan Kaufmann.

Young, F.W. & Hamer, R.M. 1987. *Multidimensional Scaling: History, Theory, and Applications*. Hillsdale, NJ: Lawrence Erlbaum Associates.

Pei, Y-F., Zhang, L., Li, J. & Deng, H-W. 2010. *Analyses and Comparison of Imputation-Based Association Methods*. Xi'an Jiaotong University, Xi'an, People's Republic of China.