# Chapter 1

# Introduction

Recent advances in molecular technology have allowed large quantities of genotypic data to be included in genetic evaluation. Two such recent advances are the single nucleotide polymorphism (SNP) chip to collect marker information and the cDNA microarray to collect gene expression information. Both of these technologies provide an opportunity to gain insight into the molecular functionality of genes and allow more accurate identification of genetically superior animals. They are high throughput devices and collect a large number of records for relatively few individuals. Many of the established statistical methods used in genetics require the number of animals to exceed the number of explanatory variables. Thus, the use of these high throughput methods demands new statistical approaches.

The aim of this thesis is to demonstrate some statistical techniques that can be applied to exploit these vast amounts of genotypic data and so enhance the accuracy of genetic selection. The thesis begins with a brief review of issues relating to genetic evaluation incorporating marker information in genetic studies and microarray technology. Emphasis is placed on the statistical aspects of the literature. A more

in-depth discussion of the relevant literature is included within each experimental chapter. The experimental chapters can be viewed as comprising two sections. The first section explores the use of SNP data to predict genetic value (Chapters 3, 4 and 5). In Chapters 3 and 4, principal component analysis (PCA) is used to reduce the dimensionality of the large predictor space of the SNPs and the principal components (PCs) of these SNPs are employed to predict breeding values. Chapter 5 compares the use of different kernels when kernel regression is applied with the aim of predicting breeding value from discrete marker information. The second section of the experimental chapters, comprising Chapters 6 and 7, examines the use of a wavelet threshold and other traditional methods to remove the nuisance spatial noise from microarray slides.

Chapters 3, 4, 5, 6 and 7 have been prepared as individual manuscripts and consequently there is some overlap within these chapters and with Chapter 2. However, some amendments have been made to their original form to aid continuity in the organization of this thesis.

# Chapter 2

# Literature Review

## 2.1 Genetic Evaluation

### 2.1.1 Introduction

Genetic evaluation is important in livestock production so that genetically superior animals can be selected from their contemporaries to be parents of the next generation. Within the last 60-70 years, planned breeding programs based on genetic principles have been implemented for a variety of species (Hammond et al., 1992, Chapter 1), accelerating the rate of genetic change. Selection has traditionally been based on phenotypic records of animals and their relatives with best linear unbiased prediction (BLUP) favoured as a tool for analysis of the data generated. Recent advances in molecular genetics, such as the discovery of genetic markers, have provided more tools for selection, which can be used to complement the traditional BLUP approach. However the inclusion of such data has required more complex statistical methods for the analysis of unbalanced data.

### 2.1.2 Best Linear Unbiased Prediction

For the last 35 years BLUP has been the preferred method of genetic evaluation in many industries. BLUP estimates have the desirable properties of having the minimum least square error in the class of linear estimators whilst being unbiased (Robinson, 1991). BLUP was largely derived by Henderson (1973) and is a general method of estimating the random effects in a mixed linear model such as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Zu} + \mathbf{e}, \tag{2.1}$$

where $\mathbf{y}$ is a vector of trait values, $\mathbf{b}$ is a vector of fixed effects with incidence matrix $\mathbf{X}$, $\mathbf{u}$ is a vector of random effects with incidence matrix $\mathbf{Z}$ and $\mathbf{e}$ is the vector of residuals such that:

$$E\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{Xb}\\ \mathbf{0}\\ \mathbf{0}\end{bmatrix}$$

and

$$\text{Var}\begin{bmatrix}\mathbf{y}\\ \mathbf{u}\\ \mathbf{e}\end{bmatrix} = \begin{bmatrix}\mathbf{V} & \mathbf{ZG} & \mathbf{R}\\ \mathbf{GZ}^T & \mathbf{G} & \mathbf{0}\\ \mathbf{R} & \mathbf{0} & \mathbf{R}\end{bmatrix}, \tag{2.2}$$

with $\mathbf{G}$ and $\mathbf{R}$ known positive definite matrices and $\mathbf{V} = \mathbf{ZGZ}^T + \mathbf{R}$.

The BLUP of $\mathbf{u}$, $\hat{\mathbf{u}}$, and best linear unbiased estimator (BLUE) of $\mathbf{b}$, $\hat{\mathbf{b}}$, are solutions to the mixed model equations:

$$\begin{bmatrix}\mathbf{X}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{X}^T\mathbf{R}^{-1}\mathbf{Z}\\ \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{X} & \mathbf{Z}^T\mathbf{R}^{-1}\mathbf{Z} + \mathbf{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\mathbf{b}}\\ \hat{\mathbf{u}}\end{bmatrix} = \begin{bmatrix}\mathbf{XR}^{-1}\mathbf{y}\\ \mathbf{Z}^T\mathbf{RZ}^{-1}\mathbf{y}\end{bmatrix}.$$

Many different forms of the linear mixed model are used to evaluate the genetic worth of animals. The basic animal model incorporates information from all relatives with or without phenotypic records to estimate breeding values (BVs). The $\mathbf{G}$ term in equation (2.2) is replaced by $\sigma_a^2\mathbf{A}$, where $\sigma_a^2$ is the additive genetic variance and $\mathbf{A}$ is the numerator relationship matrix, which describes the covariances between relatives due to the laws of inheritance. When fitting the animal model it is assumed that all base animals are sampled from the same population with mean 0 and a common variance. This assumption can be avoided by adding genetic groups to the model (Westell and Van Vleck, 1987; Westell et al., 1988; Quaas. 1988). The gametic model is used when parental BVs are the major focus and each animal's BV is expressed as the mean of its parent BVs. In addition, a Mendelian segregation term is absorbed into the error term, $\mathbf{e}$, altering the form of the covariance matrix, $\mathbf{R}$. If the contribution from the dam is not explicitly modeled in the gametic model, this is known as the sire model (Henderson, 1973).

Care must be taken when using the sire model as biased estimates may result from non-random mating due to the dam effects being omitted. Quaas et al. (1979) proposed the maternal grand sire model to allow information on the dams to be used. Features of the gametic model and animal model were combined by Quaas and Pollak (1980) to form the reduced animal model (RAM), whereby parents have their BVs estimated as in the animal model and offspring have their BVs estimating as the sum of their parents BVs plus a Mendelian segregation term. The numerator relationship matrix, $\mathbf{A}$, only considers the parents in the RAM. The reduced animal model has a computational speed advantage over the animal model because the equations of the progeny are absorbed into the equations of their parents.

#### 2.1.2.1 Multiple traits

When traits are correlated, more precise BLUPs can be obtained by fitting a multiple trait model. This model utilizes the covariance structure within traits and animals with missing values for particular traits. Selection bias can be accounted for if the traits used as a basis for selection are included in the multi-trait analysis (Sorenson and Kennedy, 1984). However, the precision of multi-trait BLUP is reliant on well estimated correlations between traits (Schaeffer, 1984).

### 2.1.3 Marker Assisted Selection

Genetic markers are polymorphisms or variations in the DNA sequence. Some of the more common marker types are mentioned here. The first DNA marker used was the restriction fragment length polymorphisms (RFLP; Elston et al. 2002; Botstein et al. 1980), which cuts DNA segments according to mutations, leaving segments of differing lengths. Hence, mutations are detected by identifying different length segments of DNA.

Another important class of markers is referred to as variable number tandem repeat (VNTR) polymorphisms, which are multiple copies of a sequence of base pairs that are repeated different number of times from allele to allele. When the number of base pairs that are repeated is small ($< 4$), this is called a microsatellite and when the number of base pairs is larger, the marker is termed a minisatellite. Most recently single nucleotide polymorphisms (SNPs) have been the marker of choice, due to their abundance (see for example, Sachidanandam et al. (2001) and Wong et al. (2004)).

Marker assisted selection (MAS) uses associations in the joint distribution of quantitative traits and marker genotypes to aid genetic evaluation. Soller and Beckmann

(1983) were among the first to discuss applications of genetic markers in selection and they identified early recognition of genetically superior animals as one potential application of MAS. Meuwissen et al. (1996) showed by simulation that the extra information obtained by using MAS could increase genetic gain by 8.8-38% initially, but this figure drops in subsequent generations. Similarly, Schaeffer (2006) concluded that MAS could be used in the dairy industry to predict very accurate EBVs of sires at birth, reducing the cost of proving bulls by 92% and increasing the rate of genetic change by a factor of 2.

Many methods have been suggested to utilize marker information to select superior animals to assist in selection. Geldermann (1975) employed a least squares procedure to estimate marker allele effects, however this method is not robust to complications, such as non-random mating, which might be experienced in the field. Fernando and Grossman (1989) demonstrated how marker information could be incorporated into a linear model to simultaneously estimate the additive effect of markers, fixed effects and the effects of all the remaining quantitative trait loci (QTL) whilst accounting for sources of bias encountered in the least squares method of Geldermann (1975). That is, equation (2.1) is re-written as:

$$\mathbf{y} = \mathbf{Xb} + \mathbf{Wv} + \mathbf{Zu} + \mathbf{e}, \tag{2.3}$$

where $\mathbf{v}$ is a vector of fixed marker effects with incidence matrix $\mathbf{W}$ and all other terms are as defined in equation (2.1). Essentially, the extra information obtained by the markers is used to modify the genetic covariances fitted to reflect the information known regarding which sections of the chromosome were inherited. Van Arendonk et al. (1994) built upon this theory to arrive at a method whereby multiple marker effects could be fitted while only a single random effect for each animal is predicted.

The recent advancement of high throughput genotyping methods has allowed many more markers to be used in studies, resulting in a change for utilizing marker information. This change in approach is required because in many instances the number of marker effects that are to be estimated is far greater than the number of data points (Lande and Thompson, 1990). Meuwissen et al. (2001) combined marker information from the whole genome to compare 3 methods of estimating the joint effect of many markers on relatively few animals, whilst not over-parameterizing the model. The first method was least squares method with only the largest marker effects included. Secondly, a BLUP method with constant variance of each segment of the chromosome was employed. The final method was a Bayesian method, whereby a prior distribution was assumed for each marker.

More recently, Gianola et al. (2006) illustrated how non-parametric methods, such as kernel regression, could be used in MAS. They suggested that parametric modeling could not handle the multiplicity of potential interactions that arise when thousands of markers are used and that non-parametric methods were an ideal alternative. A discrete binomial kernel was demonstrated by Gianola et al. (2006), however a Gaussian kernel was used in their reduced kernel Hilbert space mixed model to estimate BVs, due to its differentiability. In using the Gaussian kernel on markers, a continuous metric is imposed on the intrinsically discrete marker information and the effect of treating discrete data as continuous is unknown.

## 2.2 cDNA Microarray

### 2.2.1 Overview

Gene expression technologies are a relatively new development, which allow scientists to measure gene expression levels of thousands of genes simultaneously (Schena et al., 1995). In particular, cDNA microarrays allow measurement of transcription of a gene, a process whose function is critical in determining the phenotype of a cell. Since the technology is somewhat new, a short outline of the procedure involved in microarray experiments is given.

Nguyen et al. (2002) gave an excellent overview of basic technical and biological aspects of microarray technology. The following is a brief summary of the basic steps involved in a two channel microarray experiment:

- cDNA clones are obtained from a cDNA library, inserts are amplified and are printed to the array. The clones to be applied to the array can be chosen at random or clones believed to be of interest can be selected from a database. These clones are printed onto the array and form the spots on the arrays, which ideally would be of uniform size, shape and cDNA content. This is not the case in reality, adding to the variation of the experiment.

- A biological sample is taken from the organism of interest. RNA is isolated in the sample and cDNA is produced through reverse transcription. The cDNA is labeled with a dye, typically Cy3 (green) or Cy5 (red) that will fluoresce when excited by a laser.

- Hybridization involves two samples, each with a different dye, being applied to

the slide. Complementary cDNA binds to the clones spotted on the array, so that the amount of cDNA from the sample present on the slide and consequently the amount of each dye present is proportional to the gene expression level for each sample.

- Each slide is scanned with a laser at two different wavelengths corresponding to the excitation frequencies for Cy3 and Cy5. Two grayscale images per slide are obtained, one for each dye, reflecting the intensities of each dye, from which the expression level of each gene can be estimated. That is, each spot on the array has two intensity values, one corresponding to the sample associated with the red dye (R) and the other with the green dye (G).

The inherently complicated procedure coupled with the limit of 2 colours per slide means that many statistical challenges, such as design, normalization and analysis arise from microarray technology.

### 2.2.2 Design

The relatively high cost and the large variation associated with cDNA microarray experiments means that design is a very important issue. Fundamental considerations of classical experimental design are still applicable in the microarray context (Kerr and Churchill, 2001a; Kerr, 2003). Replication is important so that a valid estimate of error can be obtained as a contrast for the estimated effects (Fisher, 1971). Most importantly, this implies that multiple specimens should be used to sample biological material in the first phase of the experiment, so that biological variability can be estimated (McIntyre, 1955; Kerr, 2003). Other ways that replication can be incorporated into a microarray design include spotting genes multiple times on each array,

dye swaps and using the same cDNA sample on multiple arrays, all of which allow estimation of measurement error and could be considered technical replication.

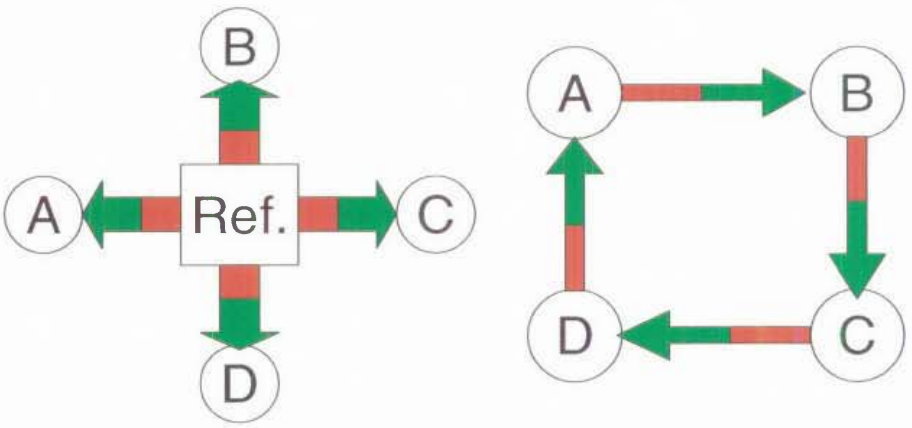
Randomization is also critical in a microarray experiment (Kerr, 2003). If the experiment is two phased, it is important that treatments are randomly allocated to animals in the first phase of the experiment. It is also important that the microarray phase of the experiment is randomized so that any technical effect that is unaccounted for can be reduced. Thus, multiple randomizations are required in a microarray experiment.

The blocking structure is particularly crucial in an experiment involving microarrays, so that where possible, effects of interest are orthogonal to each other. Thus, at first glance it appears that traditional incomplete block designs could be employed. However, Kerr (2003) indicated that blocking structure in a microarray experiment should be slightly different to a classical block design since a classical block design is typically concerned with individual treatments, whereas a microarray study would usually make inferences about the population being sampled.

Much work has been focused on deciding which biological samples to apply to each array in an experiment. It is common for graphical representations to be used to depict the allocation of samples to arrays (Kerr and Churchill, 2001a). Samples are represented by nodes or vertices and array by edges or arrows directed from the sample to be labeled with the red dye to the sample to be labeled with the green dye. Two examples of commonly discussed designs are given in Figure 2.1.

Figure 2.1: Graphical representation of a (a) Reference design (b) Loop design

(a) (b)




The reference design (Figure 2.1(a)) has been a popular choice in microarray experiments, despite its limitations. Under this design the reference sample is labeled with one dye and all other samples are labeled with another dye separately. All of the samples are compared to the reference sample, collecting a huge amount of information on the reference sample but very little on the other samples. An advantage of the reference design is that new slides and samples may be incorporated into the experiment at a later date.

The loop design (Figure 2.1(b)) is another well known design, whereby two samples can be compared via a chain of other samples. The superior efficiency of the loop design when compared to the reference design is well documented (e.g. Vinciotti et al. (2005); Yang and Speed (2002); Kerr and Churchill (2001a)).

However, the problem of microarray experimental design is rarely simple and the loop design is not necessarily the optimal design for given objectives and constraints. There are frequently constraints on the number of animals, the number of slides and

the amount of RNA that can be acquired from each sample. Similarly, the aims and required precision of different microarray experiments seldom concur. Hence it is important to be clear of the aims and constraints of an experiment when choosing the experimental design (Yang and Speed, 2002).

### 2.2.3 Normalization

Microarray experiments are prone to systematic errors such as varying RNA quality, efficiency of dye labeling and uneven washing of a sample over the array. Normalization corrects for these errors by removing some of the sources of variation that influence the measured gene expression level.

One of the most commonly used normalization tools is the locally weighted linear regression (lowess) adjustment (Cleveland, 1979; Yang et al., 2002). The ratio ('R') of *log* base 2 intensities, $log_2(R/G)$, is plotted as a function of the *log* base 2 product, ('I'), $log_2(RG)$. This is known as an 'RI' plot and sometimes an 'MA' plot. A lowess function is fitted to the plot and subtracted from the ratio to adjust for intensity dependent dye bias. Figure 6.6 on page 88 is an example of an 'RI' plot before and after lowess normalization.

Cui et al. (2003) extended the lowess approach to accommodate spatial heterogeneity by fitting the ratio of intensity as a function of the row and column position of a spot as well as the mean intensity in the lowess adjustment. However, Smyth and Speed (2003) suggested that this approach may not account for discontinuities in the spatial bias such as those encountered on the boundary of printing blocks and proposed a within print block lowess adjustment as an alternative.

Other authors have suggested modeling the nuisance spatial trends separately

to the lowess procedure. Shai et al. (2003) transformed the data to the frequency domain with the Fourier transform and applied a Gaussian filter in the frequency domain to account for nuisance spatial trends. Baird et al. (2004) proposed fitting a first order autocorrelation (AR1) structure for modeling small scale vibrations and splines for fitting global variations on a microarray following the theory established in crop science by Cullis and Gleeson (1991). Similarly, Burgueño et al. (2005) fitted an AR1 correlation structure in microarray experiments containing empty spots to account for spatial variation.

The ANOVA method (Kerr et al., 2000) models the logarithm of intensity as a function of factors such as the dye, treatment, array and interaction terms. This model was built upon by Wolfinger et al. (2001) with a mixed linear model approach and can be used to normalize microarray data.

Semi-parametric models have also been incorporated by some authors in normalizing microarray slides. Fan et al. (2004) proposed a semi-linear model that relies on spots being replicated within a slide. This approach was extended to incorporate information across slides by Fan et al. (2005) so that replication of genes within a slide was no longer required. Huang et al. (2005) proposed a two-way semi-linear model for normalization which allows uncertainty in the normalization stage to be included in the analysis of genes.

### 2.2.4 Finding Differentially Expressed Genes

One of the earliest and most basic methods employed to identify expressed genes is to examine the fold change. DeRisi et al. (1996) identified genes with $log_2(R/G)$ more extreme than $\pm$ 3 as differentially expressed. Chen et al. (1997) developed a confidence

interval for the ratio so that the significance of the ratio could be quantified. However these methods are limited to only a single array with two samples per array and rely on strong parametric assumptions. Dudoit et al. (2002) developed a permutation based t-test so that some of the parametric assumptions about the distribution of expression levels could be relaxed, while addressing the multiple testing issue. Similarly, Thomas et al. (2001) proposed a regression method which made no distributional assumptions of the microarray data (provided that sample sizes are large (Pan, 2002)), while accounting for a high false discovery rate.

Mixture models have also been employed to find differentially expressed genes. Sapir and Churchill (2000) applied an orthogonal linear regression of one colour on the other colour and modeled the residuals as a mixture of uniformly distributed differentially expressed genes and a normally distributed common component. A mixture of normal distributions was utilized by Reverter et al. (2003) to identify 3 clusters of genes, one of which was associated with differentially expressed genes. Efron et al. (2001) avoided parametric assumptions by using an empirical Bayes method, whereby the data determines the shape of the density functions. McLachlan et al. (2005) demonstrated how to used mixture models to control the false discovery rate and false negative rate in microarray studies.

Mixed linear models have also been employed to find differentially expressed genes. Kerr and Churchill (2001b) used a global ANOVA model to jointly estimate all gene expressions. Wolfinger et al. (2001) proposed a gene specific mixed model be used after normalization, but this approach has lower power than the global analysis (Wu et al., 2003). Hoeschele and Li (2005) demonstrated how the gene specific analysis could be employed in the joint linear mixed model analysis to obtain exact inference under

certain conditions. The ANOVA models of Kerr and Churchill (2001b) and Wolfinger et al. (2001) were expanded to the multivariate mixed model framework by Reverter et al. (2004) so that data from multiple studies could be analyzed simultaneously.

## 2.3 Conclusion

This brief review has examined genetic evaluation techniques and issues associated microarray studies. In Chapters 3, 4 and 5, statistical methods are presented for genetic evaluation using SNP data and these methods are a form of marker assisted selection. Chapters 6 and 7 look at modeling spatial noise in microarray experiments, which is considered a component of microarray normalization.

# Chapter 3

# Principal Components Analysis of SNP Data to Predict Breeding Value

## 3.1 Introduction

Developments in genetic technology allow large numbers of single nucleotide polymorphism (SNP) values to be scored for individuals. Within animal breeding, it is hoped that these SNPs can be used to predict the genetic merit of animals at an early stage so that superior animals can be identified for further testing or breeding.

The large numbers of SNPs that are evaluated means that the predictor variables are contained in a high dimensional space. This is referred to as the 'Curse of Dimensionality' (Bellman, 1961), a phenomenon that can be overcome by adding more animals to the experiment or by reducing the dimension of the predictor space. It

may not be practical to increase the number of animals in many cases because the required increase in the number of animals is approximately $3^{n_s}$, where $n_s$ is the number of SNPs, which can be in the thousands. Thus, it is sensible to reduce the dimension of the predictor space.

Perhaps the most widespread method used in dimension reduction is principal component analysis (PCA), which finds linear combinations amongst the multivariate responses from each experimental unit such that the variance explained is maximised. Projection pursuit is another dimension reduction technique where 'interesting' projections of the data are sought. However, the results of projection pursuit are harder to interpret than those from PCA.

Within the SNP domain, various authors have recognised that dimension reduction techniques can be used in data analysis. Dimension reduction techniques are particularly useful in SNP data since there is often a high level of interdependence between loci due to linkage. Basic linear algebra operators have been applied to eliminate aliased and linearly dependent SNPs. He et al. (2005) demonstrated how Gaussian elimination could be performed on the matrix, $\mathbf{X}$, of SNP values for all individuals so that linearly independent bases could be formed in order to find haplotype tagging SNPs. This method ensures preservation of the rank of $\mathbf{X}$ and allows very accurate full haplotype reconstruction given only the haplotype tagging SNPs. Similarly, Lin and Altman (2004) used PCA to locate SNPs that capture most of the haplotype diversity within a gene-specific region of the chromosome. They applied PCA to the matrix of SNP values, $\mathbf{X}$, and SNPs whose coefficients in the leading principal components (PCs) exceeded a threshold were identified as haplotype tagging SNPs. Horne and Camp (2004) used a similar method to infer groups of SNPs

in linkage disequilibrium (LD) so that a near optimum set of SNPs could be found to account for intragenic variation. PCA was applied to $\mathbf{X}$ and LD groups were identified by grouping together SNPs with an absolute coefficient greater than 0.4 in each significant PC. These investigations were to find associations among SNPs. This work is extended to use the information in the associations to estimate breeding value.

Analyses to predict categorical data such as disease status were developed by Hahn et al. (2003), Ritchie et al. (2001) and Ritchie et al. (2003). The technique used a non-parametric multifactor-dimensionality reduction technique to collapse multilocus information into a single dimension whilst retaining the ability to model the interaction between multiple loci. However, this method is computationally intensive, only practical for relatively few SNPs and designed for categorical data.

Meuwissen et al. (2001) developed methods for predicting breeding values (BVs) from whole genome marker maps. They concluded that selection on predicted breeding values from genetic markers could increase the rate of genetic gain. Similarly, Schaeffer (2006) discussed the implications of genome wide selection in the dairy industry and concluded that it could have a large positive effect on genetic improvement and profit.

In this chapter PCA regression is applied to genome wide SNP data with the aim of predicting genotypic merit. It extends the previous work regarding dimension reduction of markers in that information contained from SNPs taken from the entire genome is reduced to a lower dimensional space; and this information is used to predict values for a continuous response. A cross-validation method is used to select the optimal number of PCs to use in the regression, and methods to decide which PCs to include in the model are utilized to make the model more accurate. The methods

are applied to simulated and real data.

## 3.2 Materials

### 3.2.1 Real Data

These data comprised 15,380 SNPs taken from 1,546 dairy sires born between 1955 and 2001. The EBVs of each animal for milk protein percent were supplied by the Australian Dairy Herd Improvement Scheme, along with the reliability of each EBV. The EBVs are between -0.43% and 0.44% with reliabilities between 0.25 and 0.99. The mean reliability for the EBVs is 0.89.

The SNPs are recorded as 0 for the homozygote aa, 1 for the (unordered) heterozygote Aa, and 2 for the homozygote AA. Of the 23,777,480 SNP values 35.19% are aa, 25.59% Aa, 32.33% AA and 6.89% are missing values. All of these missing values are replaced with 1's, so that all of the SNP values are consistent with Mendel's first law.

### 3.2.2 Simulation

Organisms consisting of two copies of one chromosome of length 20 million base pairs were simulated. In order to simulate the effect of linkage disequilibrium (LD), a small number of chromosomes, $n_c$, were created to generate the base population. The number of founder chromosomes used was (i) $n_c = 20$ and (ii) $n_c = 200$. A total of 1,000 SNPs were placed on the chromosome, with their base pair positions randomly sampled from the integers between 1 and 20 million without replacement. Of these 1,000 SNPs, (a) $n_a = 10$, (b) $n_a = 100$ and (c) $n_a = 1{,}000$ were simulated

to have an additive effect. These effects were sampled from a Gamma distribution with shape parameter 0.59 and scale parameter 7.1 (Hayes and Goddard, 2001) and an equal probability of being positive or negative. The probability of the minor allele occuring at the $i$th site, $p_i$, was between 0 and 0.5 and randomly sampled from a uniform distribution, so that the matrix of haplotype values for the $j$th chromosome was given by:

$$B_{ij} = \begin{cases} 0 \text{ with probability } 1 - p_i \\ 1 \text{ with probability } p_i. \end{cases} \tag{3.1}$$

The first 30% of the rows of the matrix $B$ were paired up to form males and the remaining 70% paired up to form females. Random mating was performed to produce the first generation of 500 individuals. The distance between cross-overs in the breeding process was sampled from a Poisson distribution with parameter 1 million, so that each chromosome was 20 Morgans long. Thus, one long chromosome was simulated rather than many shorter chromosomes. No mutation was simulated.

The population structure was intended to be a simplified representation of the breeding structure in place in the dairy industry in Australia. The initial population of 500 animals consisted of 40 males and 460 females and random breeding was once again simulated to form a new 395 animals. Ten of these animals were assigned as new males and 385 as new females and they replaced the same number of parents in the breeding population. This process was repeated for 10 generations and the last three generations were stored.

The true molecular breeding value (MBV) for each animal recorded in the last three generations was calculated as:

$$MBV = \sum_{i=1}^{j=1000} q_i a_i, \tag{3.2}$$

and the phenotypic value as:

$$T = MBV + \epsilon, \tag{3.3}$$

where $q_i$ is the number of minor alleles (0,1 or 2) at SNP position $i$, $a_i$ is the allelic substitution effect of the $i$th polymorphism and $\epsilon$ is sampled from a N(0,$\sigma_e^2$) distribution. The predefined heritability ($h^2$) and the additive genetic variance ($\sigma_a^2$) determined $\sigma_e^2$ via the equation $h^2 = \frac{\sigma_a^2}{\sigma_a^2+\sigma_e^2}$.

## 3.3 Methods

### 3.3.1 Principal Component Analysis

Principal component analysis is a multivariate analysis technique where the aim is to reduce the dimension of a dataset comprised of many correlated variables, while still accounting for a large proportion of the variance (Jolliffe, 1986). Given a vector, $\mathbf{X}$, of random variables, the first PC is the linear function, $\mathbf{w_1^T X}$ such that var($\mathbf{w_1^T X}$) is maximised and $\mathbf{w}_1^T\mathbf{w}_1 = 1$. The $j$th PC is the linear function, $\mathbf{w}_j$, which is orthogonal to all other PCs that maximises var($\mathbf{w_j^T X}$). The problem of finding PCs is equivalent to finding the eigenvalues, $\lambda$, and eigenvectors, $\mathbf{w}$, of the covariance matrix of $\mathbf{X}$, $\Sigma$.

#### 3.3.1.1 Application of Principal Component Analysis to SNP data

The animals can be partitioned into those with observations (EBVs or phenotypes) ($K$) and those without ($U$). The animals in the set $K$ form the training set from which parameters are estimated that are to be used to predict the MBVs of the animals in the set $U$. SNPs without variation are removed from the study. The remaining SNPs are arranged into a matrix $\mathbf{X}^o = \{x_{ij}^o\}$, where $x_{ij}^o$ is the number of minor alleles (0,1

or 2) in the $i$th SNP position for the $j$th animal. PCA is performed (i) for all animals ($j \in K \cup U$) and (ii) only animals in the training set ($j \in K$).

The vector of SNP means, $\{\mathbf{x}_{i.}^{o}\}$ is computed, saved and subtracted from each column of $\mathbf{X}^{o}$ to form the matrix of $n_s$ SNPs for $n_{an}$ animals, $\mathbf{X}_{n_s \times n_{an}} = \{x_{ij}^{o} - x_{i.}^{o}\}$. Principal component analysis is performed on the matrix $\mathbf{X}$ via the expectation maximisation (EM) algorithm as described by Roweis (1997), which has an advantage in high dimensional data because it does not require computation of the sample covariance matrix.

The algorithm to find the first $n_{pc}$ PCs is:

**for** $i = 1, 2, ...., n_{pc}$ **do**

Choose a vector $\mathbf{w}_i = ({}^{1}w_i, {}^{2}w_i, ...., {}^{n_s}w_i)^T$ so that $(\mathbf{w}_i^T)\mathbf{w}_i = 1$

**loop**

(E step) Compute $\mathbf{y} = ((\mathbf{w}_i)^T(\mathbf{w}_i))^{-1}(\mathbf{w}_i)^T\mathbf{X}$

(M step) Compute $\mathbf{w}_i^{new} = \mathbf{X}\mathbf{y}^T(\mathbf{y}\mathbf{y}^T)^{-1}$

Scale $\mathbf{w}_i^{new}$ such that $(\mathbf{w}_i^{new})^T(\mathbf{w}_j^{new}) = 1$

**end loop**

Subtract the projection of each point onto the principal component from $\mathbf{X}$ to obtain $\mathbf{X}^{new}$. That is, $\mathbf{X}_j^{new} = \mathbf{X}_j - (\mathbf{w}_i \cdot \mathbf{X}_j)\mathbf{w}_i$, where $\mathbf{X}_j$ is $j$th column of $\mathbf{X}$ and '$\cdot$' denotes the dot product.

**end for**

The vector, $\mathbf{y}$, which is calculated in the E-step is the projection of the data into the one-dimensional subspace defined by the vector of weights, $\mathbf{w}^i$. The $i$th principal component is given by $\mathbf{pc}_i = (\mathbf{w}_i)^T\mathbf{X}$ and $(\mathbf{pc}_1, \mathbf{pc}_2, ..., \mathbf{pc}_{n_{pc}})$ are now ordered such

that $\mathbf{pc}_1$ accounts for the most variation in $\mathbf{X}$ and $\mathbf{pc}_{n_{pc}}$ accounts for the least variation. The principal components and rotation matrix, $\mathbf{W}_{n_s \times n_{pc}} = (\mathbf{w}_1, \mathbf{w}_2, ..., \mathbf{w}_{n_{pc}})$ are stored.

A linear model is fitted of the form:

$$\mathbf{T}_{j \in K} = \beta_1 pc_{j,1} + \beta_2 pc_{j,2} + ...\beta_{n_{pc}} pc_{j,n_{pc}} + \varepsilon, \tag{3.4}$$

where $\varepsilon \sim N(0, \sigma^2)$, $\mathbf{T}_{j \in K}$ is the phenotype of a particular trait or EBV of animal $j \in K$, $pc_{j,i}$ is the $i$th principal component for the $j$th animal and $\beta_1, \beta_2, ..., \beta_{n_{pc}}$ are the regression coefficients. This is referred to as principal component regression (PCR). To predict the MBVs of the desired animals, the estimated regression coefficients from equation (3.4) are used:

$$\hat{T}^{pred}_{j \in U} = \hat{\beta}_1 pc_{j,1} + \hat{\beta}_2 pc_{j,2} + ... + \hat{\beta}_{n_{pc}} pc_{j,n_{pc}}. \tag{3.5}$$

It is anticipated that the use of animals in the set $U$ may add noise to the PCs to be used in the PCR. In order to compare the accuracy of the PCR when PCA is performed on animals in the set $K \cup U$ to when PCA is performed on animals in the set $K$, PCA is performed on the set $K$. The regression coefficients are estimated as before (equation (3.4)). The animals whose MBVs are to be predicted are arranged into a matrix $\mathbf{Z}^o = \{z^o_{ij}\}$, where $z^o_{ij}$ is the number of minor alleles for the $i$th SNP for the $j$th animal as before. The vector of mean SNP values from the training set, $\{x^o_{i.}\}$, is subtracted from each row of $\mathbf{Z}^o$ to form the matrix $\mathbf{Z}$. The principal components are computed for these animals by:

$$\{\mathbf{pc}^T_1, \mathbf{pc}^T_2, ..., \mathbf{pc}^T_{n_{pc}}\} = \mathbf{Z}^T\mathbf{W}. \tag{3.6}$$

These PCs are used to predict the genotypic merit through equation (3.5).

#### 3.3.1.2 Supervised Principal Component Analysis

Many SNPs may have no effect on genetic merit. The inclusion of such SNPs may add noise to procedures used to predict MBVs. Supervised principal component analysis (SPCA; Bair and Tibshirani, 2004; Bair et al., 2006) is a method whereby a univariate regression is performed measuring the univariate effect of each SNP on the EBV. Only SNPs whose t-test on the regression coefficient exceed a threshold, $\theta$, are taken and PCA is performed on this subset of SNPs. This method is used for $\theta = 2$ (corresponding p-value $\approx 0.05$) and $\theta = 3$ (p-value $\approx 0.003$). The case of $\theta = 0$ is equivalent to PCA.

#### 3.3.1.3 Choosing the Number of Principal Components

Classically, methods utilizing the eigenvalues corresponding to the rows of the rotation matrix have been used in order to choose the number of PCs to keep. This includes methods such as keeping PCs with eigenvalue greater than unity, Scree plot, Horn's procedure, regression methods, Bartlett's test and the broken-stick test (see, for example Jolliffe (1986), Johnson and Wichern (1988) and Sharma (1996)). However, it has been found that such methods greatly underestimate the number of PCs needed to accurately predict genotypic merit, since not all of the important information in the SNP data is necessarily captured in the leading PCs. There is no reason why quantitative trait loci (QTL) should only occur in areas of the chromosome where there is a large amount of variability; and information of these QTL may be captured in PCs that account for a relatively small proportion of the overall variance.

A cross-validation method is used to estimate the optimal number of principal components to be used in the regression. For these real data, in order to estimate the number of PCs required, the EBVs of $n_{uk} = 150$ animals are randomly dropped from the sample and saved. These animals form the group of unknowns, $U$, and the remaining animals the group of knowns, $K$. PCR is performed and the regression coefficient are estimated with varying number of PCs being used in the regression. The MBVs of the $n_{uk}$ animals in $U$ are estimated and the correlation with their saved EBVs is examined. This process is repeated. For these simulated data, the youngest animals form the group of unknowns, $U$, and these animals have their MBVs estimated and compared to their simulated MBVs.

#### 3.3.1.4 Selection of Principal Components

Although the PCs are ordered from the PC that accounts for the most information to the PC that accounts for the least variation, this does not necessarily imply that the first PC contains the most relevant information for predicting genetic value. Thus, some of the PCs that account for a significant part of the variation of the original data may be spurious and make the linear model unsound for prediction. Three methods are used to select the PCs.

In the first method, PCs are ranked according to the proportion of variance accounted for by each PC. Secondly, the correlations are computed between each PC and the observation (EBVs or phenotypes). The PCs are ordered according to their absolute correlation with the response variable, so that the first PC fitted in the model is the most highly correlated with the response variable. The third method of ordering the PCs is a combination of the first two methods. The PCs that are most highly

correlated with the response variable may account for a very small proportion of the variation in the SNPs, making the PCR less robust. Similarly, the PCs that account for a large proportion of variance in the SNPs may not influence the observation at all. The PCs are ranked according to $|s_i|$:

$$s_i = \frac{\lambda_i \rho(\mathbf{pc}_i, \mathbf{T})}{\sum_{j=1}^{n_{pc}} \lambda_j}, \tag{3.7}$$

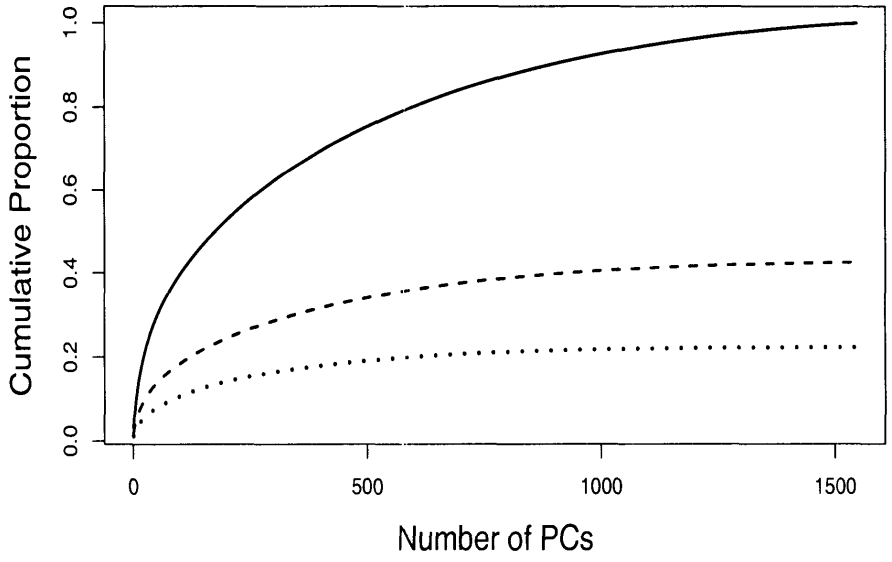
where $\lambda_i$ is the $i$th eigenvalue and $\rho(\mathbf{pc}_i, \mathbf{T})$ is the correlation between the $i$th PC and the observation.

## 3.4 Results

### 3.4.1 Real Data

Figure 3.1 displays the cumulative proportion of the variance accounted for by the PCs when PCA and SPCA are used. If all 1,546 of the PCs are taken when PCA is used, clearly all of the variance of the original data is contained. The first 200 and 500 PCs account for 50% and 75% of the variation respectively when all of the SNPs are used in the reduction. The SPCA methods do not account for 100% of the total variation when all PCs are included because not all of the original 15,380 SNPs have a t-value greater than the threshold ($\theta$). When $\theta = 2$, 42.69% of the SNPs are taken and these SNPs account for 35.54% of the total variation and when $\theta = 3$, 22.39% of the SNPs are taken which account for 18.11% of the variation in the unedited data.

Figure 3.1: The cumulative proportion of variance accounted for in multilocus allele frequencies by the PCs when: (i) PCA is used (——), (ii) SPCA is used with $\theta = 2$ (- - -), (iii) SPCA is used with $\theta = 3$ (· · · · · ·).




Pairwise plots of the EBVs of the animals and the first 3 PCs illuminate some interesting structure in the data, as displayed in Figure 3.2. The plots above the diagonal are obtained when PCA is used and plots below the diagonal are from SPCA with $\theta = 2$. Figure 3.2 distinguishes between animals born before 1995 and those born in 1995 or afterwards. This year was chosen because it divides the animals into two approximately equally sized groups. In the majority of plots above the diagonal in Figure 3.2, the year of birth of each animal influences the distribution of points. It can be seen that animals born before 1995 tend to have lower EBVs than those born in 1995 or afterwards.

When PCA is used to reduce the data, older animals tend to have a lower score for PC1 than newer animals, indicating that PC1 is in the same direction as selection

pressure. There are 2 distinct clusters in the plot of PC1 against PC2, where age defines which cluster animals belong to. A number of outliers can also be identified from the pairwise plots that arise from PCA.

When SPCA is used to reduce the data, more outliers can be identified and less variation is evident in the first four PCs. Animals of similar age are not grouped together when the PCs are plotted against each other and these plots are more elliptical in shape than their counterpart from using PCA.

Figure 3.2: Exploratory plots of the EBVs and the first 3 PCs for animals born • before 1995 and • 1995 or later. Plots above the diagonal are for the reduced data when PCA is used and below the diagonal where SPCA is used, $\theta = 2$.

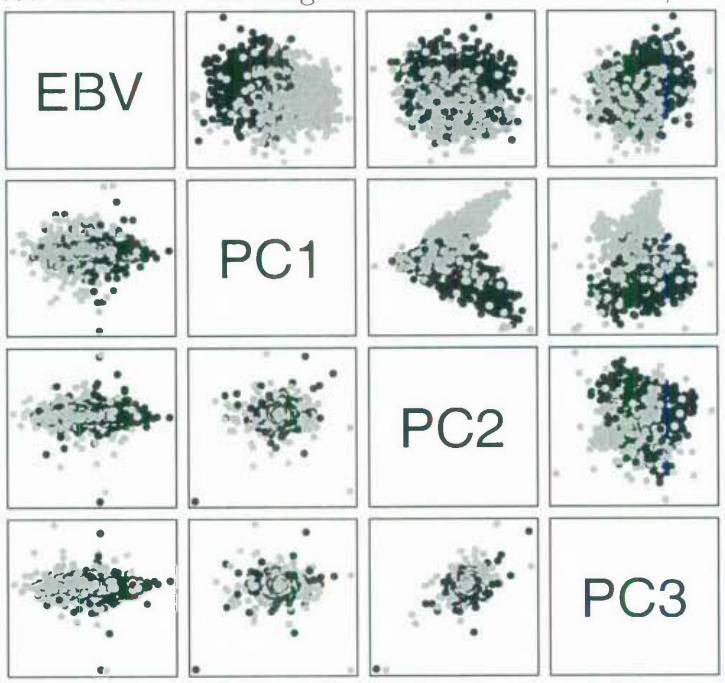
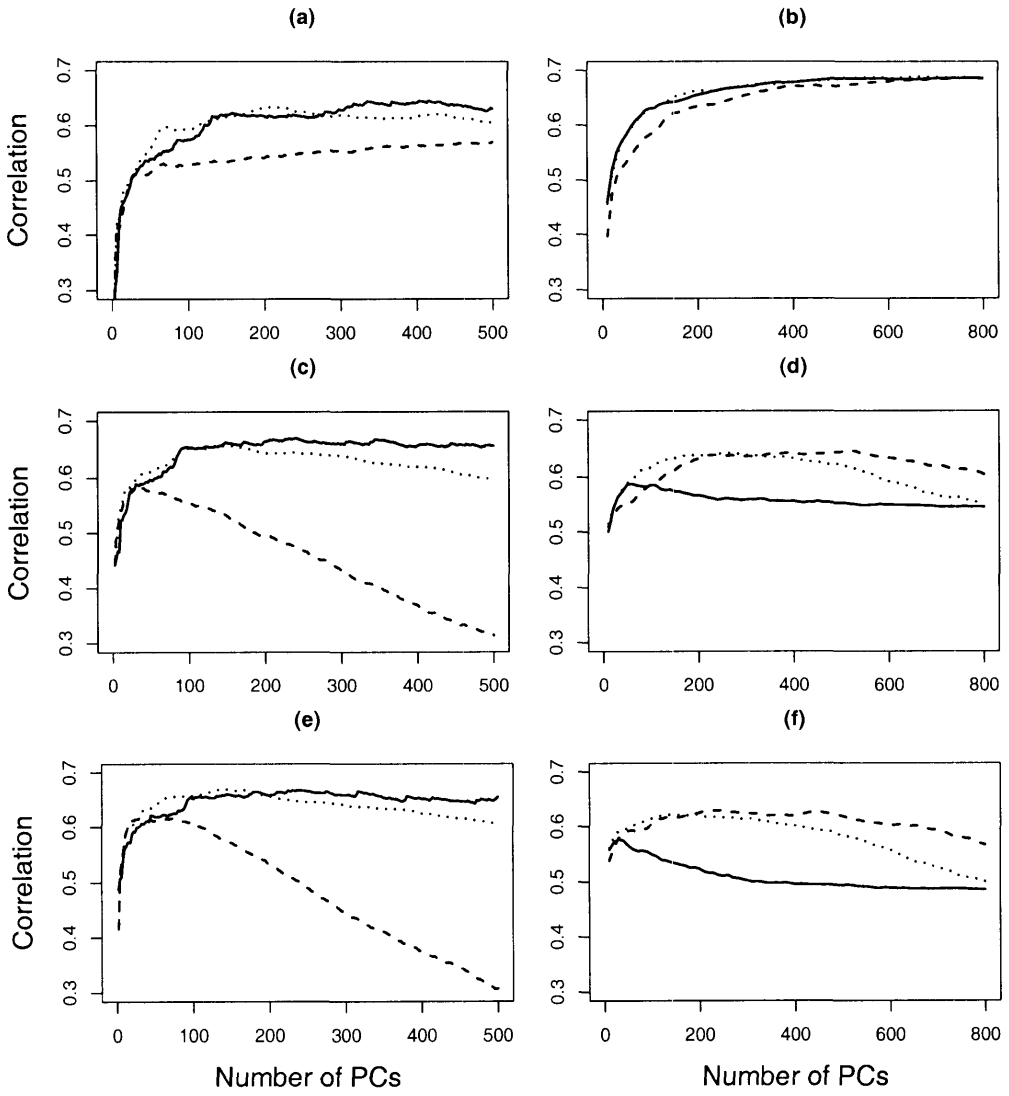


Figure 3.3 displays the mean correlation between the predicted MBVs and EBVs when the cross-validation method is repeated 40 times. When PCA is performed on all SNPs in all animals (Figure 3.3(a)) the mean correlation reaches a maximum of 0.65 when 300-500 PCs are fitted according to their eigenvalues. Before this maximum is reached the curve is not monotonically increasing, with the inclusion of some PCs in the regression reducing the predictive performance of the model. When PCs are added according to the correlation with the known EBVs, a maximum of 0.57 is obtained; and when PCs are added according to the value of $|s_i|$, the maximum is 0.63.

There is a slight improvement in predictive performance when SPCA is used on all animals (Figures 3.3(c) and (e)). This improvement is greatest for $\theta = 3$, where a maximum mean correlation of 0.67 is obtained for methods adding PCs to the regression according to $\lambda_i$ and according to $s_i$. When the correlation between the PCs and EBVs is used to determine the order that PCs are added, the maximum is reached after relatively few PCs, but then falls away quickly.

The best predictive model for these data is when PCA is performed on animals with known EBVs (Figure 3.3(b)). A maximum mean correlation of 0.69 is obtained for all three methods of adding PCs to the regression when more than 600 PCs are added. When SPCA is used only on the animals with known EBVs, the estimates are further from the known EBVs.

Figure 3.3: Correlation between estimated MBVs and EBVs with the **real** data when:
**(a)** PCA is performed on all animals ($K \cup U$) and all SNPs;
**(b)** PCA is performed only on animals with known EBVs ($K$) and all SNPs;
**(c)** SPCA is performed on all animals ($K \cup U$) and SNPs with $\theta > 2$;
**(d)** SPCA is performed only on animals with known EBVs ($K$) and SNPs with $\theta > 2$;
**(e)** SPCA is performed on all animals ($K \cup U$) and SNPs with $\theta > 3$;
**(f)** SPCA is performed only on animals with known EBVs($K$) and SNPs with $\theta > 3$.
PCs are added according to the size of the corresponding eigenvalue (——), correlation with the EBVs (- - -) and a combination of the two methods (• • • • • •) (Mean of 40 samples).

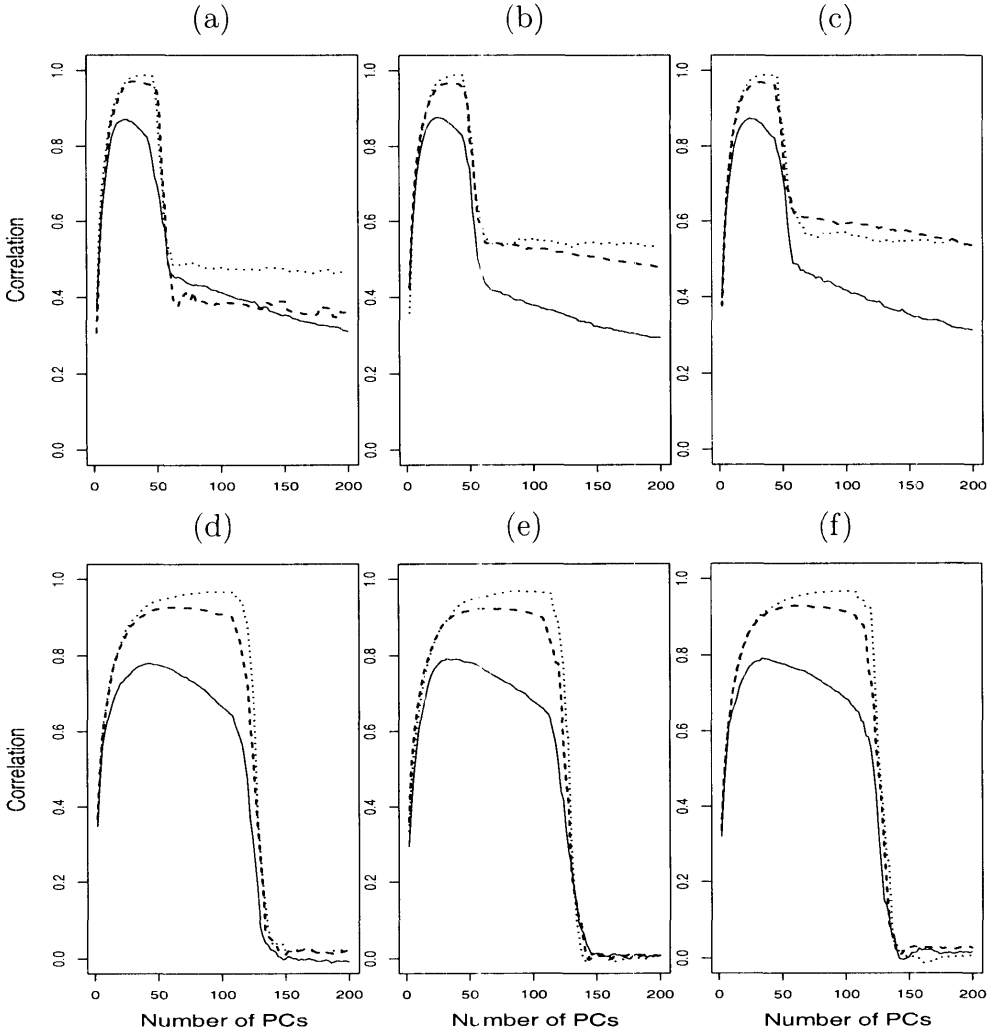


### 3.4.2 Simulation

Figure 3.4 examines the predictive performance of PCR for the simulated SNP data when the heritability, $h^2$, of the trait is varied as well as the number of SNPs with an additive effect, $n_a$, and the number of founder chromosomes in the initial population, $n_c$. The PCs are added according to the proportion of the total variation accounted for. It can be seen that the optimal number of PCs to use is about 30 for all nine combinations of $h^2$ and $n_a$ when $n_c$=20. Beyond this optimal number of SNPs, spurious PCs are fitted and the correlation between the estimated and true values decreases rapidly, before this descent becomes more gentle at about 50 PCs. As expected, the heritability of the trait influences the performance of the PCR, with higher $h^2$ values allowing better prediction of genotypic merit when the optimum number of PCs are fitted. The influence of the number of SNPs with an effect is more subtle. For low $h^2$, $n_a$ has little effect on the performance of PCR. However for $h^2 = 0.7$, and $h^2 = 0.4$ increasing the number of SNPs with an additive effect from 100 to 1000 improves the performance of PCR when more than 50 PCs are fitted.

When $n_c = 200$ the number of SNPs with an additive effect, $n_a$, has very little influence on the performance of the PCR. The $h^2$ has a larger effect when $n_c$= 200 than when $n_c$= 20, with higher $h^2$ yielding better predictive performance. More PCs are required in the regression when $n_c$=200, with around 125 PCs needed for a $h^2$ of 0.7 for optimum predictive performance.

Figure 3.4: Correlation between estimated MBV and **simulated** MBV for the unknown animals, **U** using PCR when :
**(a)** 10 SNPs have an additive effect and 20 chromosomes are in the founder population;
**(b)** 100 SNPs have an additive effect and 20 chromosomes are in the founder population;
**(c)** 1000 SNPs have an additive effect and 20 chromosomes are in the founder population;
**(d)** 10 SNPs have an additive effect and 200 chromosomes are in the founder population;
**(e)** 100 SNPs have an additive effect and 200 chromosomes are in the founder population;
**(f)** 1000 SNPs have an additive effect and 200 chromosomes are in the founder population.
Simulated heritabilities are 0.1 (——), 0.4 (- - -) and 0.7 (· · · · · ·) (Mean of 50 samples).



## 3.5 Discussion

A method has been proposed to estimate MBVs from SNP data and phenotypes, which could be EBVs or observations of a particular trait. The genome wide variation in the SNP data is used to account for the variation in EBVs by summarizing the original SNP data with PCs. The location of the SNPs is not required for the method to be used and it could be used to predict the genotypic value of animals with no phenotypic values or known relatives.

Pairwise plots of the EBVs and the first few PCs can be used as an exploratory tool for the data analyst. Such plots can be useful in identifying outliers and potentially erroneous observations. The plots can also be used as a visual representation of change in genome wide genetic make-up of a population with time or with EBV.

Far more PCs are required in the regression for these real data than were predicted for these simulated data. This may occur because of a higher number of founder chromosomes in the real data, longer chromosomes and more complicated interactions between loci in the real case. Furthermore, the maximum correlation between EBV and predicted MBV is lower for these real than for these simulated data. This is expected because these simulated data are free of genotyping errors, missing values, non-additive effects, epistasis and other potential sources of noise. We anticipate an improvement in the predictive performance of the method when animals with more reliable EBVs are used in the regression.

It has been found that, of the models considered here, the best model to estimate MBVs is the model where PCA is performed for all SNPs from only the animals in group $K$ and the rotation matrix from these animals is used to find the PCs for the animals in group $U$. The inclusion of the animals in $U$ may add noise to the PCs,

making the regression coefficients less robust.

When SPCA is performed on $K \cup U$, there is an improvement in the robustness of the PCR. Removing genes without strong effects removes some noise in the procedure. However, when SPCA is performed on $K$, the performance of the estimates reduces. In this case more information is being lost from the SNPs with small effects than the reduction in noise.

Although PCR has been shown to be robust in predicting MBV, there are some issues that need to be addressed. The use of PCA with ordinal marker data violates the assumption that PCA be used with continuous, multivariate normal data. However, it is the author's view that the high dimension of the predictor space may negate this violation.

The issue of how to handle missing values also needs to be addressed. In this study, SNPs with missing values have been assigned to be heterozygotes so that the laws of Mendelian genetics are conserved. The EM algorithm presented by Roweis (1997) can accommodate missing values, by assigning them values in the E-step, but the solution is not constrained to have a biological meaning.

Although PCA can be used with SNP data to significantly reduce the dimension of the predictor space in MBV estimation, in its current form the method requires all of the SNPs in the study to be scored. Thus, there is still a large cost associated with genotyping the animals. This cost is not as important for estimating the genetic worth of bulls in the dairy industry, where the cost of proving young bulls is many times larger than the cost of genotyping them. However, the cost of genotyping may mean that using large numbers of SNPs is not optimal. Hence, finding a smaller subset of SNPs that account for a large proportion of the genetic variation may be useful. This

is beyond the scope of this chapter, but we note the possibility of finding a smaller subset of SNPs by observing the contribution of each SNP of each PC inculded in the PCR.

The use of all SNPs in the function to predict MBV with PCR also has advantages. Where SNPs are aliased in the training set, all of these SNPs are retained in the model and PCA places equal weights on them. If all SNP effects were to be estimated individually, typically at least one of the aliased SNPs would be dropped from the model. Thus, this dimension reduction technique can account for SNPs that are completely confounded in the training set and not in the set of animlas to have their MBV estimated because of random sampling.

# Chapter 4

# Practical Considerations for Principal Component Analysis of SNP Data to Predict Breeding Value

## 4.1 Introduction

A current focus of animal breeding is to use genotypic information of animals in order to predict specified traits. Genetic markers such as single nucleotide polymorphisms (SNPs) are becoming the prevalent genotypic information. A huge amount of genetic information being is gathered, with tens of thousands of SNPs being collected and in some cases relatively few animals. This is an example of the 'Curse of Dimensionality' (Bellman, 1961), where there are many more explanatory variables than there

are observations.

One way to reduce the number of explanatory variables is to use dimension reduction techniques, such as Gaussian reduction (He et al., 2005), projection pursuit (Huber, 1985) and principal component analysis (PCA; Jolliffe, 1986). Such techniques allow much of the information in the high dimensional space of response variables to be captured within a low dimensional space. Woolaston et al. (2007) (Chapter 3) used principal component regression (PCR) on SNPs generated from a whole genome scan to predict breeding value (BV).

The impact of erroneous marker information on linkage, marker maps and finding quantitative trait loci (QTL) has been the subject of considerable research. Abecasis et al. (2001) and Douglas et al. (2000) found that relatively low rates of genotyping error could significantly reduce the power of linkage studies. Similarly Göring and Terwilliger (2000) commented that false negative QTLs may result from genotyping errors in multipoint analysis.

This chapter examines some practical considerations for using PCA on whole genome SNP scans to predict molecular breeding value (MBV). Real data are used to study the effect of the number of animals in the training set on the optimum number of principal components (PCs) to use in the PCR and the predictive performance of the PCR. Simulated data are examined to investigate the effect of erroneous SNP readings and missing SNP values on MBV estimation by PCR.

## 4.2 Materials

### 4.2.1 Data

These data are comprised of 15,380 SNPS recorded from $n_{an} = 1,546$ dairy sires. However, 298 SNPs show no variation and are removed from these data, so $n_s = 15,082$ SNPs are retained. These data are arranged into a matrix, $\mathbf{X}_{n_{an} \times n_s}$, where

$$X_{ij} = \begin{cases} 0 \text{ if the } j\text{th SNP for the } i\text{th animal is aa} \\ 1 \text{ if the } j\text{th SNP for the } i\text{th animal is aA (unordered)} \\ 2 \text{ if the } j\text{th SNP for the } i\text{th animal is AA .} \end{cases} \quad (4.1)$$

A total of 6.89% of all the SNPs were missing and were replaced with 1's to be consistent with Mendel's first law.

The dairy sires were born between 1955 and 2001 and EBVs for milk protein percent were estimated as a part of the Australian Dairy Herd Improvement Scheme. Table 4.1 gives a breakdown by years of the EBVs.

Table 4.1: Summary of the EBVs of the dairy sires.

| Year of Birth | Before 1972 | 1972-1981 | 1982-1991 | 1992-2001 | All animals |
|---|---|---|---|---|---|
| Number of animals | 10 | 82 | 457 | 997 | 1546 |
| Mean of EBVs (%) | -0.176 | -0.079 | -0.038 | 0.020 | -0.004 |
| sd(EBV) | 0.120 | 0.108 | 0.108 | 0.123 | 0.123 |
| Mean Reliability of EBVs | 0.778 | 0.940 | 0.918 | 0.874 | 0.890 |

### 4.2.2 Generating Simulated Breeding Values

The real SNP data described above were used to generate the simulated MBVs. From the 15,082 SNPs, $n_a$ were randomly sampled and taken to be quantitative trait nucleotides (QTNs). The QTNs were given additive values, $a_j$, sampled from a Gamma distribution with shape parameter 0.59 and scale parameter 7.1 (Hayes and Goddard, 2001). Each effect was randomly assigned to be positive or negative with probability 0.5 (Meuwissen et al., 2001). The simulated MBV for each animal is the sum of the additive QTN effects for that animal:

$$MBV_i = \sum_{j=1}^{n_a} X_{ij} a_j,$$

where $X_{ij}$ is defined in equation (4.1) and $a_j$ is the additive effect of an allele substitution at the $j$th QTN.

The QTNs were dropped from the set of known SNPs to investigate the case where no known SNPs are located directly on the QTN.

#### 4.2.2.1 Missing Values

The pattern of missing values in the real data was used to produce the pattern of missing data in the simulated data. A summary of the pattern of missing values is displayed in Table 4.2. The number of animals with missing values for the $j$th SNP in the real data, $mv_j$ is saved. Each simulated SNP is randomly assigned to a $mv_j$, with $mv_j$ animals, picked at random, recorded as having a missing value for the simulated SNP.

Table 4.2: Distribution of the number of missing SNP values.

| Number of missing values | 0-100 | 101-200 | 201-300 | 301-400 | 400-1400 | 1400-1546 |
|---|---|---|---|---|---|---|
| % of SNPs | 94.42 | 3.08 | 0.72 | 0.13 | 0.01 | 1.64 |

#### 4.2.2.2 Genotype Errors

A stochastic-error model (Akey et al., 2001) wass used to simulate the pattern of mis-typing. The probability of erroneously observing a homozygote, $aa$, given the true genotype is the (unordered) heterozygote $Aa$, $P(aa|Aa) = p$. Similarly, $P(Aa|aa) = p$ and $P(aa|AA) = p^2$. The value of $p$ was chosen such that the overall error rate is 1% ($p = 0.007$), 2% ($p = 0.015$), 5% ($p = 0.038$), 10% ($p = 0.074$) or 20% ($p = 0.143$).

## 4.3 Methods

### 4.3.1 Principal Component Regression

These data are divided into the training set, $K$, comprised of animals with known EBVs and the set of animals whose MBVs are to be predicted, $U$. PCA is performed on the covariance matrix of $X_{ij}$ for $i \in K$, and the projection matrix of eigenvalues, $\mathbf{W}_{n_p \times n_s}$, is saved. That is, PCA is performed on the SNPs independently of the EBVs and phenotypes. The PCs of the SNPs in the set $K$ are used as explanatory variables and the EBVs are the response variables in multiple linear regression. The SNPs in the set $U$ are projected into the principal subspace by multiplication with the projection matrix, $\mathbf{W}$. These projections are used in the multiple linear regression equation to predict MBV.

#### 4.3.1.1 Order of Principal Components

The order that the PCs are added to the PCR equation can influence the predictive performance of the regression when a given number of PCs are included in the PCR (Woolaston et al., 2007, or Chapter 3). Three methods of determining the order in which PCs are added to the regression are considered here.

**Method 1:** PCs are added according to the proportion of variance for which they account. The first PC added is the PC that accounts for the most variation in these data. This method may not be ideal because the first PC does not necessarily contain the most information regarding genetic merit and may add noise to the regression.

**Method 2:** PCs are added according to their absolute correlation with the EBV. The first PC added is the PC whose absolute correlation with EBV is the highest. The first PCs added with this method could only account for a small amount of variation in the SNPs and such PCs can be inaccurate, adding further noise to the regression.

**Method 3:** A compromise between Method 1 and Method 2. PCs are added to the PCR according to proportion of overall variance accounted for, corrected by the correlation of the PC with the MBV. This ensures that the first PCs added account for a large proportion of variance, so that they are stable and that they are correlated with EBV. PCs are added according to their value of $|s_i|$, where:

$$s_i = \frac{\lambda_i \rho(\mathbf{pc}_i, \mathbf{EBV})}{\sum_{j=1}^{n_{pc}} \lambda_j} \tag{4.2}$$

and $\lambda_i$ is the eigenvalue corresponding to the $i$th PC, $\mathbf{pc}_i$ is the $i$th PC, **EBV** is the vector of EBVs and $n_{pc}$ is the number of PCs. The first PC added to the PCR has

the largest value of $|s_i|$.

### 4.3.2 Examining Predictive Performance with Number of Animals

These real data are divided into the training set, $K$, comprised of animals with known EBVs and the set of animals whose MBVs are to be predicted, $U$. The number of animals in the training set $K$, $n_k$, should influence the predictive performance of the PCR. This influence is investigated empirically by varying the number of animals in the training set and examining the correlation between the EBVs and estimated MBVs of animals in the set $U$. The best correlation when $j$ animals are in the set $K$, $a_j$, is calculated as:

$$a_j = \max_{j \in (1,2,\ldots,n_k-1)} \frac{1}{n_r} \Sigma_{i=1}^{n_r} \rho(\mathbf{EBV}_i, \mathbf{T}_{ij}^{pred}),$$

where $\rho$ is the correlation function, $\mathbf{EBV}_i$ is the EBVs of the animals in $U$ for the $i$th iteration, $\mathbf{T}_{ij}^{pred}$ are the predicted MBVs for animals in $U$, for the model with $j$ PCs, $n_r$ is the number of replicates and $n_k$ is the number of animals in the training set, $K$. Similarly, the optimal number of PCs, $j$, to include in the model is investigated as the number of animals in the training set is varied. The optimal number of PCs when $j$ animals are in the training set, $p_j$, is calculated as:

$$p_j = \arg\max_{j \in (1,2,\ldots,n_k-1)} \frac{1}{n_r} \Sigma_{i=1}^{n_r} \rho(\mathbf{EBV}_i, \mathbf{T}_{ij}^{pred}).$$

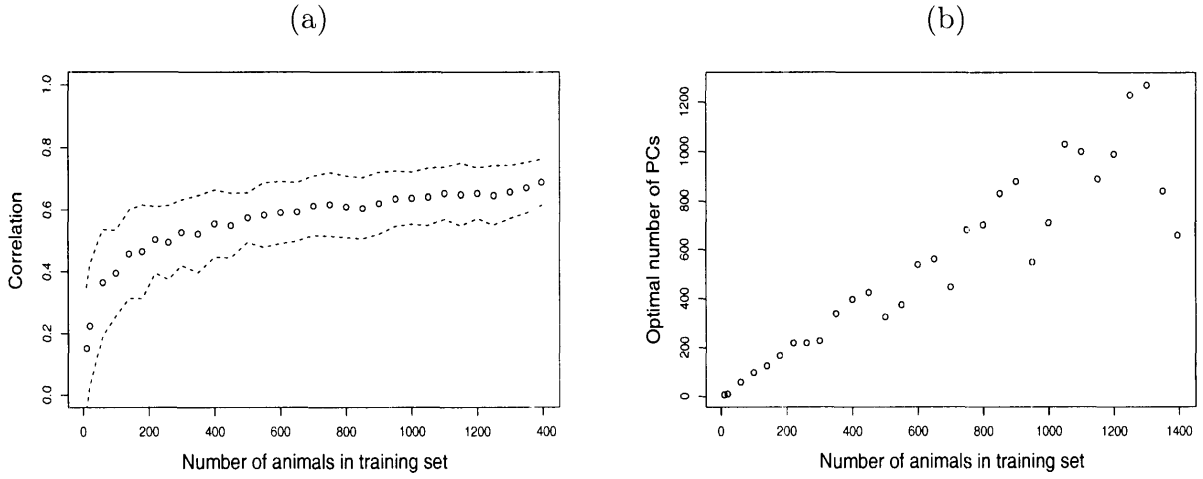
## 4.4 Results

### 4.4.1 Number of Animals

Figure 4.1 (a) shows the effect of increasing the number of individuals with EBVs on the predictive performance of the model after the cross validation method is used on these real data. Increasing the number of animals in the training set improves the correlation between EBVs and estimated MBVs, with this increase in correlation diminishing as more animals are used in the training set.

Figure 4.1 (b) suggests a linear relationship between the optimal number of PCs to include in the multiple regression and the number of animals in the training set. However there is increasing variance in the optimal number of PCs for larger $n_k$. The optimal number of PCs to fit is only slightly less than the number of animals in the training set for most values of $n_k$, but still ensures about 200 degrees of freedom for error.

Figure 4.1: The effect of the number of animals in the training set on (a) the predictive performance of PCR (95% confidence interval shown), (b) the optimal number of PCs in the PCR (Mean of 50 replications for each point). There were 150 animals in the set of unknown animals, U, in each case.




### 4.4.2 Ordering of Principal Components

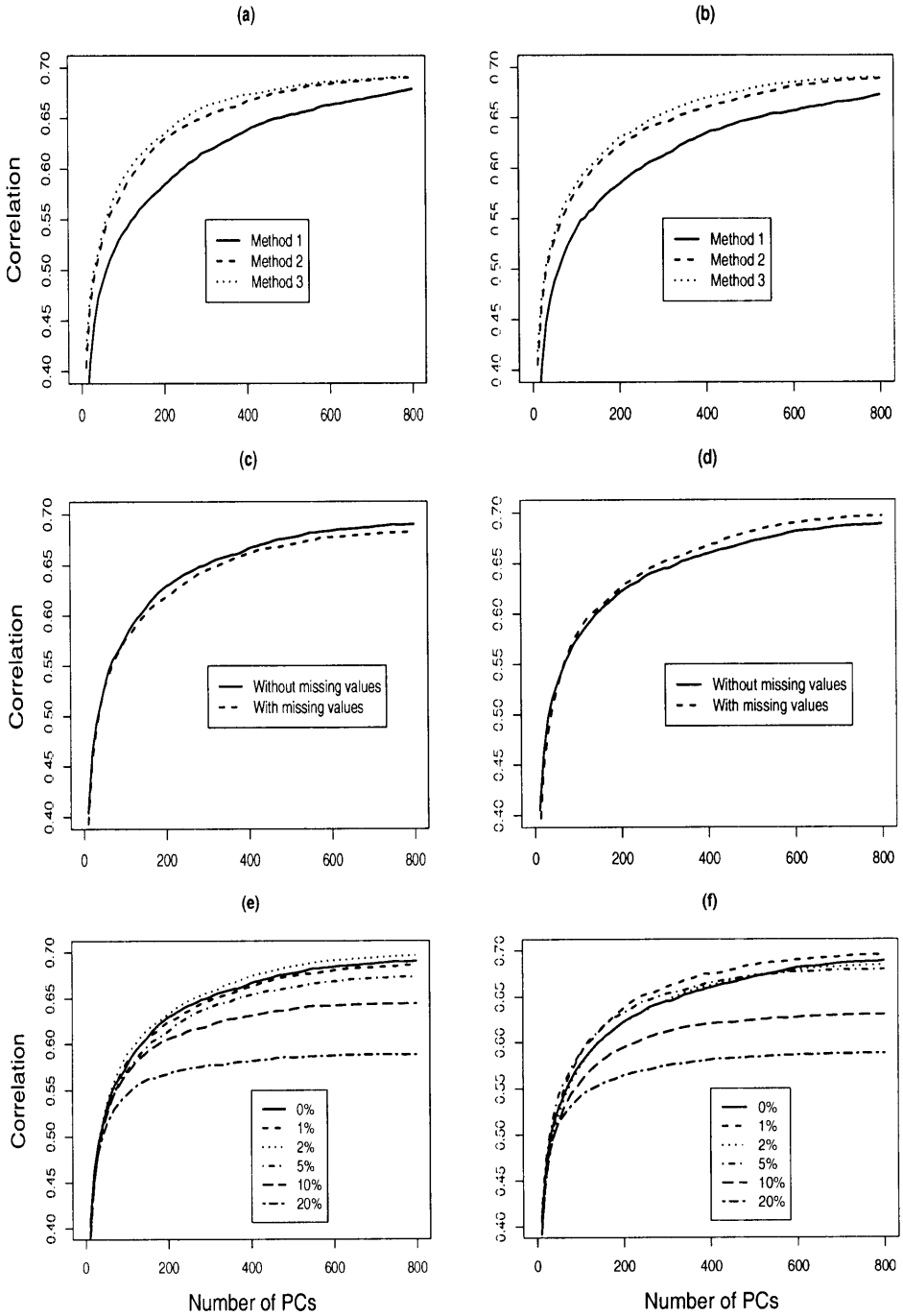
Figures 4.2 (a) and (b) show the mean predictive performance of the three different methods for adding PCs to the PCR for the simulated data for 50 replicates. Method 3, which adds the PCs according to their value of $|s_i|$ (equation (4.2)), is the best when both $n_a = 100$ and $n_a = 1000$ SNPs have an additive effect. Method 1, which adds PCs according to the proportion of variation in the SNPs they account for performs the worst. Method 2, which adds PCs according to their correlation with the EBV performs only slight worse than method 3.

### 4.4.3 Missing Values

Figures 4.2 (c) and (d) illustrate the effect of missing SNP data for these simulated data, with the mean correlation of 50 replicates being presented for Method 1. When

$n_a = 100$ SNPs have an additive effect, it is slightly detrimental to have missing values in the data. Conversely, when $n_a = 1000$ SNPs have an additive effect, it appears slightly beneficial to have missing values in the data. However, the difference in predictive performance of PCR for the case where all SNPs are accurately known and where there are missing values is small and probably due to random sampling.

Figure 4.2: Correlation between estimated MBV and simulated MBV when:
**(a & b)** Three methods for adding PCs to the PCR are compared for 100 and 1000 SNPs with a simulated additive effect respectively;
**(c & d)**Method 1: Missing data is simulated for 100 and 1000 SNPs with a simulated additive effect respectively.;
**(e & f)** Method 1: Genotyping error is simulated for 100 and 1000 SNPs with a simulated additive effect respectively. (Mean of 50 samples).



### 4.4.4 Genotyping Errors

Figures 4.2 (e and f) and Table 4.3 show the effect of erroneous SNP values on the predictive performance Method 1. Once again, the mean correlation of 50 replicates is presented. The error rate has a similar effect when $n_a = 100$ and when $n_a = 1000$, with an increase in error rate reducing the performance of the model. However, an error rate of between 0% and 2% has little effect on the correlation between the known simulated MBV and the estimated MBV. A 5% error rate has a small effect on the correlation. When a 10% error rate is present in the SNP data, there is a large reduction in the accuracy in predicting MBV.

Table 4.3: The maximum mean correlation (standard error) between the simulated MBVs and estimated MBVs for varying SNP error rates (50 replicates).

| % Error | 0 | 1 | 2 | 5 | 10 | 20 |
|---|---|---|---|---|---|---|
| $n_a = 100$ | 0.690 (0.005) | 0.686 (0.005) | 0.696 (0.005) | 0.673 (0.006) | 0.644 (0.006) | 0.588 (0.007) |
| $n_a = 1000$ | 0.688 (0.005) | 0.695 (0.005) | 0.684 (0.005) | 0.679 (0.005) | 0.631 (0.006) | 0.588 (0.007) |

## 4.5 Discussion

The relationship between the optimal number of PCs to include in the regression and the number of animals in the training set, $n_t$, suggests that almost all of the PCs contain information relevant to predicting MBV. However, the increase in variance in the optimal number of PCs to use with an increase in $n_t$ indicates that while the last PCs to be added to the regression still contain some information useful in predicting MBV, these PCs also contain a significant level of noise.

An extrapolation of Figure 4.1(a) would imply that significant improvements could be made in predicting MBVs by increasing the number of animals with already predicted EBVs in the training set. Similarly, it is anticipated that an increase in the reliability of the EBVs of the animals in the training set would further improve the accuracy of predicted MBVs.

Woolaston et al. (2007) found that adding PCs according to the value of $s_i$ was only slightly better than adding PCs according to $\lambda_i$ and significantly better than adding PCs according to their correlation with EBV for the real data described in this chapter. However, for the simulated data examined here adding PCs to the regression according to $\lambda_i$ is significantly inferior to the other two methods. This would indicate that the PCs with smaller eigenvalues are less robust to complications experienced in real data such as missing values, genotyping error and epistasis.

Missing values and genotyping errors may not have a large influence when PCA is used to predict MBVs because the method does not place a great deal of weight on each SNP. Hence, if a particular SNP is measured incorrectly, the information from surrounding SNPs helps to reduce the impact of the erroneous SNP. Roweis (1997) suggested using an EM algorithm that could accommodate missing values when using PCA. However, the results here would indicate that assuming that all missing SNP values are heterozygotes is an acceptable way of accounting for missing SNP values when PCA is used to predict MBV.