STATISTICAL METHODS TO INTERPRET GENOTYPIC DATA

By
ALEX WOOLASTON

A THESIS SUBMITTED FOR THE DEGREE OF DOCTOR OF PHILOSOPHY
OF THE UNIVERSITY OF NEW ENGLAND
March, 2007
DECLARATION

I certify that the substance of this thesis has not already been submitted for any degree and is not currently being submitted for any other degree.

I certify that to the best of my knowledge any help received in preparing this thesis, and all sources used, have been acknowledged in this thesis.
Acknowledgements

First and foremost, I would like to thank my supervisors, Dr. Bruce Tier and Dr. Robert Murison for their inspiration, guidance and for always being available to advise me. To Bruce, thanks for your patience and enthusiasm. To Bob, thank you for inspiring me to study statistics and encouraging me every step of the way.

I would like to acknowledge financial support from the Animal Genetics and Breeding Unit (AGBU) as well as the CRC for Innovative Dairy Products. In addition, thank you to everyone at AGBU, the School of Rural Science and Agriculture and the School of Mathematics, Statistics and Computer Science who contributed to the valuable discussions at various seminars and group meetings.

To my office mates, Khusro, Matias and Rob, thanks for your support and help. Thanks also to my ‘2nd office’ mates, Brad, Aaron, Cedric, Boyd and Lindsay for your friendship and motivation.

A big thank you must go to my parents for their endless encouragement and also a special thanks to Dad for his advice in putting together this work. Finally, thanks to Katie for her support.
Abstract

Recent developments in genetic techniques have provided high throughput tools such as single nucleotide polymorphism (SNP) chips and cDNA microarrays to assist in genetic selection. Such high throughput devices necessitate new statistical approaches so that the massive amounts of data gathered can be exploited in an effective manner. This thesis describes some statistical methods that can be applied to SNP data and microarray data.

Firstly, the use of SNP data to predict molecular breeding value (MBV) is studied. Principal component analysis (PCA) is used to summarize the variation of the high dimensional SNP space within a smaller dimensional projection space of principal components (PCs). It is demonstrated how the PCs can be used in principal component regression (PCR) to predict the MBV of dairy cattle from their SNP values alone with both simulated and real data. Highly reliable estimated breeding values (EBVs) are available for the real animals. A cross-validation method is used to predict MBVs for dairy sires, with a correlation of 0.69 between the EBVs and estimated MBVs obtained for these real data. The impact of erroneous SNP values, missing SNP values and the number of animals with known EBVs genotyped is also examined. Through simulation, it is found that erroneous SNP values of greater than 2% reduce the accuracy of prediction, whereas the number of missing SNP values has
little impact on the accuracy of prediction. As expected, an increase in the number of animals with already known EBVs increases the accuracy of prediction.

Kernel regression is used to predict MBV from the intrinsically discrete SNP data. Binomial kernels, which treat the SNP values as a discrete variable, and a Gaussian kernel, which imposes a continuous structure on the marker data, are employed and compared. It is empirically demonstrated that the Gaussian kernel outperforms the binomial kernel when used in Nadaraya-Watson kernel regression.

Secondly, statistical methods to account for the nuisance spatial trends found in microarray slides are assessed. Wavelets are proposed as a method of modeling spatial effects in two colour cDNA microarrays where the spatial error component may be represented as a fractal surface. This method is compared with smoothing splines plus first order autoregressive detrending using data collected from mice in a time-course experiment. Two schemes for selecting control genes are also assessed for these data, (i) pre-determined and (ii) the genes that do not over- or under-express throughout the experiment. It is shown that the spatial adjustment and the set of control genes can influence the interpretation of test genes. Results from this microarray study are also used to generate simulated data to assess the models to remove spatial trends. The wavelets threshold approach is the most successful when the nuisance spatial trends in the images are rough and fractal, but there is little difference between the models for images with smoother spatial bias.
## Contents

**Acknowledgements**  iii  
**Abstract**  iv  

**1 Introduction**  1  

**2 Literature Review**  3  

<table>
<thead>
<tr>
<th>Section</th>
<th>Pages</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1 Genetic Evaluation</td>
<td>3</td>
</tr>
<tr>
<td>2.1.1 Introduction</td>
<td>3</td>
</tr>
<tr>
<td>2.1.2 Best Linear Unbiased Prediction</td>
<td>4</td>
</tr>
<tr>
<td>2.1.2.1 Multiple traits</td>
<td>6</td>
</tr>
<tr>
<td>2.1.3 Marker Assisted Selection</td>
<td>6</td>
</tr>
<tr>
<td>2.2 cDNA Microarray</td>
<td>9</td>
</tr>
<tr>
<td>2.2.1 Overview</td>
<td>9</td>
</tr>
<tr>
<td>2.2.2 Design</td>
<td>10</td>
</tr>
<tr>
<td>2.2.3 Normalization</td>
<td>13</td>
</tr>
<tr>
<td>2.2.4 Finding Differentially Expressed Genes</td>
<td>14</td>
</tr>
<tr>
<td>2.3 Conclusion</td>
<td>16</td>
</tr>
</tbody>
</table>
3 Principal Components Analysis of SNP Data to Predict Breeding Value

3.1 Introduction ................................................. 17
3.2 Materials .................................................. 20
  3.2.1 Real Data ............................................... 20
  3.2.2 Simulation ............................................. 20
3.3 Methods ................................................... 22
  3.3.1 Principal Component Analysis ......................... 22
    3.3.1.1 Application of Principal Component Analysis to SNP data ........................................... 22
    3.3.1.2 Supervised Principal Component Analysis ........... 25
    3.3.1.3 Choosing the Number of Principal Components ... 25
    3.3.1.4 Selection of Principal Components ................. 26
3.4 Results ................................................... 27
  3.4.1 Real Data ............................................... 27
  3.4.2 Simulation ............................................. 32
3.5 Discussion ............................................... 34

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

4.1 Introduction ................................................. 37
4.2 Materials .................................................. 39
  4.2.1 Data .................................................... 39
  4.2.2 Generating Simulated Breeding Values ................ 40
    4.2.2.1 Missing Values .................................... 40
CONTENTS

4.2.2.2 Genotype Errors ........................................ 41
4.3 Methods .................................................. 41
4.3.1 Principal Component Regression .......................... 41
4.3.1.1 Order of Principal Components ....................... 42
4.3.2 Examining Predictive Performance with Number of Animals . 43
4.4 Results .................................................. 44
4.4.1 Number of Animals ....................................... 44
4.4.2 Ordering of Principal Components ....................... 45
4.4.3 Missing Values ......................................... 45
4.4.4 Genotyping Errors ....................................... 48
4.5 Discussion ............................................... 48

5 Kernel Regression of SNP Data to Predict Breeding Value 50
5.1 Introduction ............................................... 50
5.2 Materials ................................................ 52
5.2.1 Real Data ............................................... 52
5.2.2 Simulated Data ......................................... 53
5.3 Kernel Regression ......................................... 55
5.3.1 Binomial Kernel ......................................... 56
5.3.1.1 Missing Data ......................................... 57
5.3.2 Gaussian Kernel ......................................... 58
5.3.3 Predicting MBV of Each Animal From All Other Animals . 59
5.3.4 Predicting a Group of Animals From All Other Animals . 59
5.4 Results .................................................. 60
5.4.1 Simulated Data ......................................... 60
CONTENTS

5.4.2 Real Data .................................................. 63
  5.4.2.1 Predicting the MBV for Each Animal ............ 63
  5.4.2.2 Predicting the MBV for a Group of Animals .... 65
5.5 Discussion .................................................. 66

6 Spatial Effects in cDNA Microarray Slides ................. 68
  6.1 Introduction .............................................. 68
  6.2 Materials and Methods .................................... 71
    6.2.1 Data .................................................. 71
      6.2.1.1 First Phase of the Experiment ................. 71
      6.2.1.2 Second Phase of the Experiment ............... 72
    6.2.2 Statistical Models ................................... 73
      6.2.2.1 Spatial Mixed Linear Model .................... 74
      6.2.2.2 Wavelets ........................................ 76
    6.2.3 The Sample Variogram ................................ 78
    6.2.4 Control Genes ......................................... 78
    6.2.5 Analysis of Variance ................................ 79
    6.2.6 Separability and Fractals ............................ 79
    6.2.7 Calculating Fractal Dimension ....................... 81
      6.2.7.1 Box Counting Estimator ......................... 81
      6.2.7.2 Variogram-based Estimator ...................... 82
  6.3 Results .................................................. 82
    6.3.1 Results of Spatial Models ........................... 82
    6.3.2 Control Genes ......................................... 84
    6.3.3 Fractal Dimension .................................... 86
CONTENTS

A 126

A.1 Algorithm for simulation of population and SNPs .... 126
A.2 Accuracy of kernel regression estimate .............. 130
List of Tables

4.1 Summary of the EBVs of the dairy sires. ........................................ 39
4.2 Distribution of the number of missing SNP values. ......................... 41
4.3 The maximum mean correlation (standard error) between the simu-
lated MBVs and estimated MBVs for varying SNP error rates (50
replicates). ......................................................................................... 48
5.1 Distribution of the percentage of missing SNP values per animal. ... 52
5.2 Summary of the EBVs of the dairy sires. ........................................ 53
5.3 Mean accuracies (standard error) of prediction and mean optimal smooth-
ing parameters for binomial kernel regression (Kernel 1) and Gaussian
kernel regression (Kernel 2) for these simulated data (100 replications
for each simulation). ........................................................................... 62
5.4 Accuracy of predicting the MBV of 150 animals randomly sampled
from the remaining 1,396 animals (100 replicates). ......................... 65
7.1 Standard deviation of the difference between simulated and estimated
intensity reading. The simulated background noise is found with (a)
the AR1 × Splines model and (b) the Wavelets method from a previous
experiment. All standard errors were 0.01 or less. .......................... 102
7.2 Mean computation time (seconds) for fitting a spatial model for one slide. The times given for the simulated spatial noise are for Gaussian fields with a fractal dimension of 2.2. 106
List of Figures

2.1 Graphical representation of a (a) Reference design (b) Loop design. 12

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 $\theta = 3$ (••••••). 28

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, $\theta = 2$. 29
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 (\(\cdots\)), correlation with the EBVs (\(\cdots\)) and a combination of the two methods (\(\cdots\cdots\cdots\cdots\)) (Mean of 40 samples).
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).

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.
LIST OF FIGURES

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).

5.1 Accuracy of prediction for real data for varying smoothing parameter, \(k\): (a) Binomial kernels (b) Gaussian kernel.

6.1 Representation of the experimental design.

6.2 Decomposition of the grand error term, \(E\), for red, slide 15 with the (a) Splines, (b) AR1 x Splines and (c) Wavelets models.

6.3 Sample variograms for red, slide 15 with (a) No spatial trend modeled. (b) Splines. (c) AR1 x Splines. (d) Wavelets.

6.4 (a) Location of control genes found with the Wavelets model. (b) Location of the pre-specified control genes. (c) Contour plot of the spatial trend for red, slide 15 using the control genes identified using the Wavelets model. (d) Contour plot of the spatial trend for red, slide 15 using the pre-determined control genes.

6.5 Distributions of fractal dimension estimates of all 15 slides using (a) The fdim package for the red images. (b) The fdim package for the green images. (c) Box counting, red images. (d) Box counting, green images. (e) Variogram-based estimator, red images. (f) Variogram-based estimator, green images.
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.6</td>
<td>RI plots for slide 1 (a) before the lowess correction and (b) after the lowess correction.</td>
</tr>
<tr>
<td>6.7</td>
<td>Pre-determined control genes. Data-determined control genes. AR1 x Splines. Wavelets. (a) Interaction of Gene NM_007414 with challenge one, strain AJ. (b) Interaction of Gene AB041540 with challenge two, strain C57BL6.</td>
</tr>
<tr>
<td>6.8</td>
<td>State space diagram for each challenge (1 and 2) and strain (A and B) combination for the 64 genes that are identified as genes of interest after the Wavelets and AR1 x Splines models are used to remove spatial trends. A line indicates that the gene is over- or under-expressed at the corresponding time point.</td>
</tr>
<tr>
<td>7.1</td>
<td>Simulated intensities of the genes.</td>
</tr>
<tr>
<td>7.2</td>
<td>The layout of printing blocks and printed spots for the simulated array.</td>
</tr>
<tr>
<td>7.3</td>
<td>Standard deviation of the difference between simulated and estimated intensity reading. The spatial noise is simulated with Gaussian fields with (a) stable and Cauchy correlation functions and (b) stable correlation function.</td>
</tr>
<tr>
<td>7.4</td>
<td>Empirical cumulative distribution function of the estimated value minus the simulated value for a simulation with stable and Cauchy correlation functions and with fractal dimension 2.2.</td>
</tr>
</tbody>
</table>