Marker-based prediction of hybrid maize performance using genetic evaluation data

ing. Steven Maenhout

Thesis submitted in fulfillment of the requirements for the degree of Doctor (PhD) in Applied Biological Sciences

Supervisors:

- Prof. dr. Bernard De Baets
  Department of Applied Mathematics, Biometrics and Process Control, Faculty of Bioscience Engineering, Ghent University; Ghent, Belgium
- Prof. dr. ir. Geert Haesaert
  Department of Biosciences and Landscape Architecture, University College Ghent; Ghent, Belgium
- Prof. dr. ir. Erik Van Bockstaele
  Department of Plant Production, Faculty of Bioscience Engineering, Ghent University; Ghent, Belgium

Members of the examination committee:

- dr. Alain Charcosset
  Equipe Génétique Quantitative et Méthodologie de la Sélection, Unité Mixte de Recherche de Génétique Végétale; Gif-sur-Yvette, France
- ir. Bruno Claustres
  Unité de recherche R2n, RAGT Semences; Rodez, France
- Prof. dr. ir. Marnik Vuylsteke
  Department of Plant Systems Biology, Flanders Institute for Biotechnology, Ghent University; Ghent, Belgium
- Prof. dr. ir. Dirk Reheul
  Department of Plant Production, Faculty of Bioscience Engineering, Ghent University; Ghent, Belgium
- Prof. dr. Godelieve Gheysen (Secretary)
  Department of Molecular Biotechnology, Faculty of Bioscience Engineering, Ghent University; Ghent, Belgium
- Prof. dr. ir. Norbert De Kimpe (Chairman)
  Department of Organic Chemistry, Faculty of Bioscience Engineering, Ghent University; Ghent, Belgium

Dean:

- Prof. dr. ir. Guido Van Huylenbroeck

Rector:

- Prof. dr. Paul Van Cauwenberge

Marker-based prediction of hybrid maize performance using genetic evaluation data

ing. Steven Maenhout

Thesis submitted in fulfillment of the requirements for the degree of
Doctor (PhD) in Applied Biological Sciences

Please refer to this work as follows:
Maenhout, S. (2010). Marker-based prediction of hybrid maize performance using genetic evaluation data. PhD thesis, Ghent University, Ghent, Belgium.

ISBN-number: 978-90-5989-364-1

Voorwoord

Meer dan acht jaar geleden diende ik mijn ontslag in bij mijn toenmalige werkgever om als voltijds student een predoctorale opleiding te volgen. Het idee om als industrieel ingenieur een doctoraat te maken was slechts enkele maanden voordien ontstaan als een mogelijke uitweg voor de geestdodende cultuur die eigen bleek te zijn aan de sector waarin ik verzeild was geraakt. De enige motivatie voor deze stap was de utopische gedachte dat ik de tijd zou krijgen om mijn eigen ideeën uit te werken. Hoewel er op dat moment totaal geen sprake was van een mogelijke financiering voor dit plan, kreeg het de onvoorwaardelijke steun van mijn ouders, waarvoor ik hen uitermate dankbaar ben. Zij waren de eersten in een lange rij van mensen die mij hebben geholpen en gesteund bij de realisering van dit doctoraat.

Ik wil mijn dankbaarheid uiten tegenover mijn promotor Prof. dr. ir. Erik Van Bockstaele, die mij na het afronden van mijn predoctorale opleiding heeft opgevangen in het toenmalige Departement voor Plantengenetica en -veredeling (DvP) van het huidige ILVO. Zonder deze ingreep zou het ganse project een vroegtijdige dood gestorven zijn. Mijn jaar op het DvP was vruchtbaar in meerdere opzichten. Onder toezicht van mijn begeleider dr. ir. Jan De Riek werd het basisidee van deze doctoraatsstudie voor het eerst op papier gezet. De gerenommeerde expertise van het DvP liet mij ook toe om contact te leggen met de wetenschappelijke kern van het Franse veredelingsbedrijf RAGT R2n. De meeste componenten van deze doctoraatsstudie maken dan ook gebruik van gegevens die werden aangebracht door dit bedrijf. Deze bron van informatie zou echter volledig onbeheersbaar zijn geweest zonder de oprechte wetenschappelijke interesse en openheid van de RAGT medewerkers, waar de talloze bijdrages van Thierry Bouhet, Bruno Lefèvre, Bruno Claustres en Michel Romestant een specifieke vermelding verdienen.

Mijn eerste contact met mijn promotor Prof. dr. Bernard De Baets dateert ook uit de voornoemde DvP periode. Ik ben tot de dag van vandaag vereerd dat hij bereid was mij te begeleiden doorheen de wiskundige aspecten van dit werk. De bijdrage van Bernard omschrijven als louter wiskundig doet de waarheid trouwens onrecht aan daar hij eveneens

fungeerde als een bron van inspiratie, motivatie en zelfvertrouwen. Bovendien was hij de drijfveer voor de verschillende wetenschappelijke publicaties en presentaties die voortvloeiden uit onze samenwerking.
Ondanks het vertrouwen van mijn beide promotoren bleek het vinden van een financieringsbron voor het uitwerken van mijn idee nog steeds een onoverkomelijke hindernis. Gelukkig kon ik beginnen als academiseringsassistent op mijn 'oude school' waardoor het pad naar dit doctoraat eindelijk zichtbaar leek te worden. Onder de begeleiding van Prof. dr. ir. Geert Haesaert kreeg ik daar het vertrouwen en de vrijheid die al zo lang op mijn verlanglijstje stonden. Bovendien profiteerde mijn werk aanzienlijk van zijn kennis van plantenveredeling en genetica en kon ik hem ongebreideld lastig vallen met vragen en problemen van allerlei aard. Het weinige dat ik daar kan tegenover stellen is een oprechte dankbetuiging in het naar alle waarschijnlijkheid meest gelezen deel van mijn doctoraat.
De laatste maar meest belangrijke persoon die hier een bedanking verdient is mijn geliefde Chantal. Nog meer dan iedereen anders heeft ze mij gesteund en geholpen doorheen dit ganse traject van ontslag tot promotie. Jouw gedrevenheid in het leven is voor mij een continue motivatie geweest om dit doctoraat tot een goed einde te brengen.

Table of Contents

List of Figures

List of Tables

List of Symbols and Notations

This section lists the symbols and acronyms that will be used in this dissertation.

Acronyms

AFLP	amplified fragment length polymorphism
AIC	Akaike's information criterion
AIS	alikeness in state
AR1	first-order autoregressive
BLP	best linear prediction
BLUP	best linear unbiased prediction
BNO	coancestry estimator (Bernardo, 1993)
CD	generalised coefficient of determination
CD_{min}	minimum required generalised coefficient of determination
CoC	coefficient of coancestry
CS	compound symmetry
ECD	empirical cumulative distribution
GCA	general combining ability
G×E	genotype-by-environment interaction
IBD	identical by descent
ISSS	Iowa Stiff Stalk Synthetic
JAC	Jacard similarity measure
KKT	Karush-Kuhn-Tucker
LD	linkage disequilibrium
LOI	coancestry estimator (Loiselle and Graham, 1995)
MET	multi-environment trial

MLE	maximum likelihood coancestry estimator (Thompson, 1975)
OPV	open pollinated variety
PED	pedigree-based coefficient of coancestry
PEV	prediction error variance
PIC	polymorphism information content
PSD	positive semi-definite
QP	quadratic programming
REML	restricted maximum likelihood
RKHS	reproducing kernel Hilbert space
RMSE	root mean squared error
SCA	specific combining ability
SMO	sequential minimal optimisation
SNP	single nucleotide polymorphism
SSR	simple sequence repeat
SVM	support vector machine
ε-SVR	ε-insensitive support vector machine regression
VC	Vapnik-Chervonenkis
WAIS	weighted alikeness in state

Scalar, Vector and Matrix Notations

x	scalar
$\hat{x}$	estimated value for x
$\|x\|$	absolute value of x
$\boldsymbol{x}$	vector
$\langle \boldsymbol{x}, \boldsymbol{y} \rangle$	dot-product between vectors $\boldsymbol{x}$ and $\boldsymbol{y}$
$\|\boldsymbol{x}\|$	norm of a vector, generally referring to the Euclidean norm
$\boldsymbol{X}$	matrix
$\boldsymbol{X}'$	transpose of matrix $\boldsymbol{X}$
$\boldsymbol{X}^{-1}$	inverse of matrix $\boldsymbol{X}$
$\boldsymbol{X} \otimes \boldsymbol{Y}$	the Kronecker product between matrices $\boldsymbol{X}$ and $\boldsymbol{Y}$
$\boldsymbol{I}$	identity matrix

Symbols

$\mathbb{N}$	the set of natural numbers
$\mathbb{R}$	the set of real numbers
$\mathbb{C}$	the set of complex numbers
$\mathbb{R}^n$	set of real-valued vectors of size n
$f() = f(\cdot)$	function
$f(x)$	evaluation of function f at x
$\|f\|$	norm of function f, generally referring to the 2-norm
$\partial_{x_i} f(\boldsymbol{x})$	partial derivative of $f(\boldsymbol{x})$ w.r.t. the i-th component of $\boldsymbol{x}$
$\mathbb{R}^{\mathbb{R}^n}$	set of functions from $\mathbb{R}^n \to \mathbb{R}$
$K(\cdot, \cdot)$	kernel function
$I(i, i')$	indicator function, equals 1 if $i = i'$ and 0 otherwise
L	Lagrangian
$\mathcal{H}$	Hilbert space
$\mathcal{X}$	input space
$\mathcal{Y}$	output space
$\mathcal{F}$	hypothesis space
$\mathcal{D}_m$	sample of size m
$O()$	(worst case) order of an algorithm
$\mathrm{P}(A)$	probability of event A
$\mathrm{P}(A \mid B)$	conditional probability of event A given B
$\overset{\text{ibd}}{=}$	allele identity by descent from a common ancestor
$\overset{\text{ais}}{=}$	allele identity by state
$\overset{\cancel{\text{ibd}}}{=}$	allele identity but not by descent from a common ancestor
h^2	heritability
R^2	squared Pearson correlation, coefficient of determination

CHAPTER 1

Introduction, objectives and outline

1.1 Introduction

Predicting the properties of the potentially unconceived offspring of two future parents seems like a daunting task. On the other hand, the resemblance between parents and their progeny has always inspired farmers and breeders to iteratively select the most promising genotypes as parents for establishing the next generation. In fact, this process has had such a major impact on the agronomic level of production of today's domesticated crop and animal species, that it is often difficult, if not impossible, to trace them back to their ancestral origins. Therefore, these crude forms of phenotypic selection can be interpreted as the first attempts towards progeny prediction. The accuracy of these implicit prediction models is rather hard to assess as it depends largely on the heritability of the trait under study, which in turn depends partially on the selection process itself.

The quest for accurate prediction models that can assess the agronomic potential of future offspring is driven almost exclusively by economic reasons. If time and budget were unconstrained, one would just perform the cross between the two candidate parents and test the qualities of their progeny. In terms of today's plant and animal breeding programmes, this would mean that thousands if not millions of new parental crosses need to be made with subsequent phenotypic testing of their resulting progeny. In practice, this trial-and-error process has been replaced with intelligent breeding schemes that are built on the knowledge of quantitative genetics that has been accumulated in the last 100 years. In this respect, the development of hybrid varieties in several agronomically important crop species, can be considered as one of the seminal accomplishments in all of agricultural science. Maize (*Zea mays* L.) is the unrivalled showpiece of this successful breeding approach and can be considered to be one of the most important crops known to mankind. The world-wide area

designated to maize production is estimated to be more than 142 million hectares with a total estimated yearly production of 637 million ton.

The prediction of the agronomic performance of single-cross maize hybrids has always been an extremely active topic of research. The expected financial returns of a reliable prediction model for hybrid maize is an important but not entirely exclusive reason for this focus. Maize plants can be selfed for several generations which allows for the creation of nearly completely homozygous inbred lines. Such an inbred line always creates gametes with an identical allele configuration over all genes. This means that all progeny that is obtained by crossing two inbred lines is genetically identical. This genetic uniformity obviously has several advantages. The uncertainty surrounding the unknown allelic configurations in the gametes provided by the parents, can be taken out of the equation. Furthermore, as long as both inbred lines are available, their unique and uniform offspring can always be recreated and retested, allowing for very accurate phenotypic characterisations. These arguments have turned maize into the *de facto* model plant for studies involving progeny prediction. The advent of doubled haploid technology has further raised the stakes, as it allows to skip the time-consuming process of inbreeding. This implies that most of the breeding programme's budget is now drained by phenotypic testing of the newly created hybrids, exactly the part we are hoping to replace by accurate genomic prediction models.

Over the last 30 years, the field of biotechnology has evolved at an incredible pace. The current state of molecular marker technology allows plant breeders to obtain dense molecular fingerprints of their material with modest budgetary consequences. Furthermore, the complete genome sequence is already available for several plant and animal species and next-generation sequencing technology platforms are expected to vastly reduce the current sequencing efforts. The developments in the field of computer science are equally impressive, providing hardware and software platforms that can collect, analyse and store vast amounts of information in seconds. Advanced machine learning algorithms are capable of detecting relevant patterns in data streams which would overwhelm most classic statistical approaches.

All these observations lead towards the main question to be answered in this dissertation: is it possible, using the currently available arsenal of molecular and computational tools, to make accurate predictions on the agronomic performance of candidate single-cross maize hybrids?

1.2 Research objectives

The primary goal of this dissertation is to develop a methodology that allows to assess the agronomic potential of maize hybrids, based on the molecular fingerprints of their parents. This prediction model needs to be constructed by detecting patterns in the available phenotypic data that is collected during the routine genetic evaluation trials of the hybrid maize breeding programme of the private company RAGT R2n. However, it should be stressed that the detected patterns themselves, presumably in the form of marker trait associations, are considered to be of lesser importance. To meet this primary goal, several research objectives need to be defined.

The formulation of the primary goal does not specify a particular modelling approach. The use of genetic evaluation data, however, limits the option list considerably. The problem lies in the inherent unbalanced nature of this kind of data. Different hybrids are tested in different environmental circumstances, which makes it very hard to estimate a single agronomic score that allows to compare the genetic potential of these hybrids on the same scale. The linear mixed model framework does allow to calculate Best Linear Unbiased Predictors (BLUPs) of breeding values from unbalanced phenotypic data. Unfortunately, linear mixed modelling does not solve all problems. The breeding programme of RAGT R2n has produced phenotypic measurements on thousands of hybrids, while the available budget for genotyping only allows to obtain a limited set of marker scores. This means that we can either genotype a large set of inbred lines, using only a limited selection of molecular markers, or that we can genotype only a few important inbred lines, using a very dense, genome-covering molecular fingerprint. If we succeed in finding the optimal trade-off between these opposing criteria, we also need a way to actually identify the optimal selection of hybrids, having an exact and predefined number of parental inbred lines. In a similar way, we need to identify a set of molecular markers with fixed cardinality such that the resulting genome coverage is maximised. Finding a solution to these problems constitutes as the first research objective.

BLUPs, expressing the agronomic potential of the selected hybrids, can be obtained by fitting an appropriate linear mixed model to the available phenotypic data. In these models, it is commonly assumed that the genotypic components of the hybrids can be fitted as random effects for which the covariance model can be structured as some function of the pairwise coefficients of coancestry. The required coancestry estimates can be obtained from pedigree or marker information. Unfortunately, the marker-based estimation procedures that are described in scientific literature do not guarantee that the resulting coancestry

matrix will be at least positive semi-definite (PSD), a mathematical requirement for using it as a covariance model in a linear mixed model analysis. Furthermore, several procedures allow to obtain estimates that do not lie within the unit interval, contradicting the original definition of the coefficient of coancestry and hampering a straightforward interpretation of the estimated variance components. The second research objective is therefore to develop a marker-based coancestry estimation procedure that is PSD, always produces estimates within the unit interval and is specifically designed for use in hybrid breeding programmes. The molecular fingerprints of the parental inbred lines provide a very high dimensional predictor space in which each hybrid represents a specific point. We want to learn how to express the genetic potential of these hybrids as a function of their predictors (i.e. their specific position in the predictor space). This function should not be restricted to be linear and should be able to fit higher level interactions between the predictors, if these are relevant. Unfortunately, the high-dimensionality of the predictor space represents a major problem for most classic regression approaches. In this respect, kernel-based methods, a relatively recent development in the field of machine learning, might very well provide an escape from this curse of dimensionality. In particular, the group of support vector machine-based methods allows to fit non-linear classification and regression functions to very high dimensional data, while at the same time, the risk of overfitting is minimised and the resulting function allows for an efficient and fast evaluation of new hybrids. The third research objective is therefore to explore and optimise the use of the support vector machine regression framework, generally denoted as ε-insensitive support vector machine regression (ε-SVR), for the construction of a hybrid prediction model.

The accuracy of a prediction model is quite often assessed by means of some cross-validation procedure. This approach is scientifically justified if the set of training examples is sufficiently large. However, most end-users of such a prediction model mistrust this form of accuracy determination, generally claiming that the results are positively biased. In this respect, accuracy measures of hybrid prediction models are mistrusted even more, as environmental conditions tend to have a great impact on the agronomic performance that is actually measured on the field. As these environmental conditions are difficult to control, predicted and actual field measurements can deviate quite substantially, depending on the heritability of the trait under study. The fourth research objective therefore aims to evaluate the usefulness of ε-SVR-based hybrid prediction models from the perspective of the end-user, which is basically the maize breeder. This means that we need to determine how good the ε-SVR-based predictions correlate with actual field measurements taken in a validation trial. Moreover, these results need to be compared with those of more conventional prediction approaches.

1.3 Research outline

The structure of the dissertation follows the logical order that is imposed by the four research objectives: data selection, coancestry estimation, ε-SVR prediction model optimisation and model validation. In reality however, these different research parts were not necessarily performed in this particular order. The data selection approaches presented in Chapter 5 for example, were needed to select the empirical data that is used in the following chapters. In reality, the development of these procedures was a long and difficult journey which was completed long after the phase in research in which they were actually required. This means that the described selection of empirical data is most likely not optimal with respect to the accuracy of the resulting prediction models. A similar story is at the bottom of the developed coancestry estimation procedure described in Chapter 6. Its delayed finalisation explains why the linear mixed model formulations in later chapters do not make use of this estimator for modelling the covariance between hybrids. Notwithstanding these minor inconsistencies in the presented time line, it was opted to structure this dissertation according to the logical instead of the chronological order of the presented research components.

Chapter 2 introduces the topics of heterosis, hybrid breeding and prediction of agronomic performance. This chapter starts by describing the historical developments that have paved the road towards the commercial exploitation of heterosis in maize. Subsequently, an overview of the most commonly adhered theories on the molecular foundations of heterosis is provided. The chapter is concluded by an overview of earlier attempts to predict the agronomic performance of maize hybrids, giving a good idea of the challenges that lie ahead.

The mathematical foundations of the support vector machines framework is the topic of Chapter 3. First, the concept and relevant properties of reproducing kernel Hilbert spaces are introduced. Next, the focus is shifted towards the topic of structural risk minimisation, a statistical learning framework which forms the foundations of ε-SVR. This machine learning technique is gently introduced by restricting the set of candidate functions to be linear. The following section describes how this approach can be extended to result in non-linear ε-SVR models by applying the kernel trick. The chapter concludes by detailing the sequential minimal optimisation procedure, a commonly used technique to solve the quadratic optimisation problem that lies at the heart of ε-SVR.

Chapter 4 gives a description of the empirical data that was used in the different research components that are presented in this dissertation. Some limited details on the genetic

structure of the maize breeding pool and the available phenotypic data from the private breeding company RAGT R2n are provided. Next, the ad-hoc rules which have guided the selection of genotyped inbred lines are discussed. This selection of inbred lines has been fingerprinted by means of SSR and AFLP markers for which details are provided at the end of this chapter.

Chapter 5 deals exclusively with data selection issues. A framework is presented that allows to select a fixed number of inbred lines from a large set of unbalanced phenotypic data such that the resulting genomic prediction model has a superior prediction accuracy. This problem setting is approached by combining theoretical results on data connectivity issues with specific algorithms from the field of graph theory. A similar framework allows to identify the maximal genome covering subset of markers of fixed cardinality. This chapter concludes by demonstrating how one can identify the optimal trade-off between the number of genotyped inbred lines and the size of the molecular fingerprint, by means of a simulation study.

Chapter 6 introduces the newly developed weighted alikeness in state or WAIS coancestry estimator. First, the derivation and mathematical proof of the PSD property are provided. Next, the concept of matrix bending is introduced and a newly developed MCMC-based bending procedure is discussed. The suitability of WAIS for modelling the covariance between hybrids in a linear mixed model analysis is compared to that of commonly used marker-based coancestry estimation using both simulated and actual maize breeding data.

Chapter 7 explores the suitability of the ε-SVR framework for predicting the agronomic performance of maize hybrids based on the molecular fingerprints of their parents. The unbalanced nature of the selected phenotypic data is tackled by means of a linear mixed model analysis. The covariance between genetic components is modelled by means of coancestry estimates obtained from SSR, AFLP or pedigree information. The impact of the choice of kernel function on the accuracy of the ε-SVR prediction models is discussed.

Chapter 8 further optimises and validates the prediction framework established in Chapter 7. A finetuning of the linear mixed model analysis allows to improve the prediction accuracy of the subsequent ε-SVR models considerably. The robustness of ε-SVR is compared to that of a competing prediction method by reducing the number of training examples or size of the molecular fingerprint. Prediction accuracy is estimated by means of various cross-validation schemes and compared with the results of a validation field trial.

Chapter 9 is the final chapter of this dissertation. The initial research objectives are confronted with the results and conclusions that were gathered during the different stages of the presented research. This scientific reflection is finalised by a discussion of expected developments and prospects in the field of genomic selection.

CHAPTER 2

Heterosis and hybrid prediction

2.1 History of hybrid maize

'Nature thus tells us, in the most emphatic manner, that she abhors perpetual self-fertilisation ... For may we not infer as probable, in accordance with the belief of the vast majority of the breeders of our domestic productions, that marriage between near relations is likewise in some way injurious, that some unknown great good is derived from the union of individuals which have been kept distinct for many generations?'

Charles Darwin, 1862

In these words, Charles Darwin concludes his book on the Fertilisation of Orchids. It is tempting to introduce this quote as the official beginning of heterosis research but in truth, many examples of the beneficial effects of outbreeding were already well documented by that time. The increased strength and endurance of the mule for example, was already noted as far as 4000 years ago. Likewise, the detrimental effects of inbreeding were probably known even before that time as incest was forbidden or at least frowned upon in most historically important societies, ancient Greek and Egyptian cultures and some of the European royal families being notable exceptions to this rule. Despite the tenure of the aforementioned conclusion, Darwin did not recognise the beneficial effects of outbreeding as being opposite to the detrimental effects of inbreeding. He did however, inspire other scientists like dr. William J. Beal, who made report of a field experiment where hybrid maize (*Zea mays* L.) seeds were produced by sowing two different maize varieties in alternating rows and detasseling the seed-bearing (i.e. female) variety. The resultant hybrid maize seeds produced plants with 53% higher yields than either parent (Beal, 1878). Several

other experiments confirmed these results but it was not until the landmark paper of Shull (1908) that the roles of both inbreeding and outcrossing in the exploitation of heterosis were clarified. He proposed a new breeding strategy that abandoned the Darwinian principle of inbreeding avoidance and relied on cycles of explicit self-fertilisation followed by a controlled cross-pollination of carefully selected purebred lines.

Although the principles of Shull were sound, the low seed yield of the parental inbred lines prevented the adoption of this promising strategy by maize breeders. This problem was overcome by the double-cross method proposed by East and Jones (1919), at the cost of a slight reduction in vigour and uniformity of the resulting four-way hybrids. As most of these developments took place in the United States, hybrid varieties steadily replaced the open-pollinated varieties in the US corn belt. In 1930, hybrids represent only 1% of the total US maize acreage, but this proportion increases rapidly towards 50% in 1940. By the 1950s, the turnover is complete and the great bulk of maize throughout the United States is hybrid (Khanna, 1991). This commercial success motivates the hybrid's conquest of the European continent where maize hybrids are being developed by crossing the corn belt dent inbred lines with lines developed from well-adapted European open-pollinated varieties (OPV).

Despite the tremendous success of hybrid maize varieties, the relationship between genetic distance and combining ability is not well understood. Second and third-cycle hybrids are being developed from the self-fertilisation of elite inbred line crosses. The idea is to select combinations of inbred lines that compensate each other's weaknesses and little effort is put into the preservation of genetic distance or variability. The reinvention of single-cross hybrids requires the concept of heterotic patterns which slowly crystallises in the late 1960s and early 1970s (Tracey and Chandler, 1988). Inbred lines are partitioned into groups according to their population of origin and demonstrated combining abilities when crossed with inbred lines belonging to other groups. Iterative cycles of crossing, self-fertilisation and selection of lines belonging to the same heterotic group assures the preservation of the genetic distance between inbred lines belonging to different heterotic groups. Hybrids for which both parents belong to distinct heterotic groups are therefore more likely to exhibit a positive heterosis effect for yield, the most important trait from a commercial perspective. This system should also preserve the genetic variability at the population level (Lu and Bernardo, 2001) but due to the short-term breeding goals of the seed companies, the overwhelming majority of the inbred lines trace their pedigree back to only two OPVs, Reid Yellow Dent and to a far lesser extent, Lancaster Surecrop (Lee and Tracey, 2009). For example, the last available report on the genetic diversity of US hybrid maize indicates that 88% of the maize seed produced in 1984 included germplasm derived

from Reid Yellow Dent (Darrah and Zuber, 1986). Currently, the debate is still ongoing whether or not further genetic gain can be expected from the narrow genetic variability that is manifested in today's commercial breeding pools.

2.2 Genetic basis of heterosis

The term heterosis was introduced by Shull (1914) as a shorthand for 'stimulation of heterozygosis'. In one of his later papers, Shull (1948) clarified that the word was purely descriptive and was not intended to imply any genetic cause or explanation. He also provided a more explicit definition of heterosis as 'the increased vigour, size, fruitfulness, speed of development and resistance to diseases and pests manifested in crossbred organisms as compared with corresponding inbreds as a specific result of unlikeness of the constitution of the uniting parental gametes'.

Despite the great economical impact of hybrid varieties, the genetic basis of heterosis is today as much a matter of debate as it was in the early days of its discovery. The amount of literature devoted to the topic is daunting to say the least, with contrasting opinions and conclusions being published almost on a weekly basis. From the beginning, two non-exclusive hypotheses have been suggested namely, the dominance and the overdominance theories.

The dominance theory was first proposed by Davenport (1908) and rests on the assumption that the dominant genes in both parents complement each other in their hybrid offspring, masking each other's recessive, deleterious alleles. There were two main objections against this hypothesis. (1) If the dominance theory holds, it should be possible to stack these beneficial dominant alleles in one superior inbred line. As nobody has ever managed to create such a high yielding, homozygous inbred line, the initial hypothesis must be false. (2) If the dominance theory holds, the distribution of F_2 phenotypes should be skewed as the occurrence of the dominant genotypes over all loci is expected to follow a binomial distribution with a success probability of $\frac{3}{4}$. As the observed distribution of F_2 phenotypes is generally symmetric, the dominance theory must be false. Both objections are however easily overthrown by noting that the number of loci influencing a quantitative trait like yield, is generally very large which (1) makes the odds of fixing the dominant allele at each of these loci of a single inbred line extremely small and (2) removes the skewness of the binomial distribution as it asymptotically approximates the normal distribution (Collins, 1921).

The competing overdominance theory traces back to Shull (1908, 1911) and East (1908).

The idea is that the heterozygous state of a gene influencing a heterotic trait is more advantageous than the homozygous states of either of the two alleles. In other words, the two distinct alleles of the heterozygote both give a positive and nearly additive contribution to the observed phenotype, while a homozygous inbred line has to manage with only one of these allelic effects. The near additivity is hypothesised to originate from the pleiotropic divergence of the allelic functions (East, 1936). The overdominance theory did not find much acceptance as there were hardly any convincing examples of overdominant loci. Jones (1917) also indicated that the identification of a true overdominant locus is generally problematic as its beneficial allele might be tightly linked to a deleterious recessive allele of another gene, a state generally referred to as pseudo-overdominance. Furthermore, if the overdominance theory holds, it should be impossible to improve the performance of inbred lines and this turned out to be quite feasible. As a consequence, until the middle to late 1940s, the dominance hypothesis was generally accepted (Crow, 2000).

A paper of Hull (1945) tipped the scale over to overdominance. The paper contained little truly persuasive arguments but was embraced by many plant breeders as they were confronted with the failure of mass selection to substantially increase yield, which was somewhat unexpected according to the dominance theory. Nevertheless, most breeders opted for a breeding scheme that was shown to be effective under both theories, the reciprocal recurrent selection advocated by Comstock and Robinson (1952), today still the most commonly used scheme in hybrid breeding programmes. It employs the well-known concepts of the general combining ability (GCA) and specific combining ability (SCA) that were introduced earlier by Sprague and Tatum (1948).

The time window in which the overdominance theory held sway was short, as Sprague and Russell (1956) reported their results on what was called the definitive experiment. Over many years, two maize populations were each selected for increased yield in the hybrids produced by crossing them with a separate but identical inbred tester line. Under the overdominance theory, hybrids that originate by crossing inbred lines belonging to the two populations are expected to show no heterosis as both have been accumulating the same alleles to complement the tester line. By contrast, these F_1 hybrids showed increased yield, which confirmed the dominance theory.

Recently, the development of new molecular tools has given a new stimulus to heterosis research. A plethora of studies have analysed heterosis-associated gene expression in a multitude of species by comparing expression patterns of selected genes in inbred lines and hybrids or by performing high-throughput gene expression analyses via microarray profiling or GeneCalling (Hochholdinger and Hoecker, 2007). The results of these stud-

ies are contradictive and confusing to say the least, as several empirical proofs of each candidate theory has been published, included the early disregarded theory of epistasis. As a personal opinion, I think the statement of Sprague (1983) still captures our level of knowledge on the subject: "Studies have shown that additive and dominance gene effects are generally much greater than other types of gene effects. Additive effects are precisely those which respond to selection. Specially designed experiments have shown that both overdominance and epistasis exist, but neither has been shown to be important at the population level. ... Thus, as far as the maize breeder is concerned, a pragmatic solution to the dominance-overdominance controversy has been reached. Additive and dominance effects provide a satisfactory model for the heterosis and for the rather remarkable progress achieved through breeding."

2.3 Hybrid prediction

The reciprocal recurrent selection scheme used by hybrid maize breeders requires the development of a large number of new inbred lines on a yearly basis. These new inbred lines, belonging to a particular heterotic group, tend to have a good general combining ability when crossed with inbred lines belonging to a complementary heterotic group, as decades of recombination and selection have adapted both gene pools to combine well with each other. However, to become a commercial success, a hybrid, in addition to these pedigree-derived beneficial additive effects, generally needs to manifest a considerable, positive heterosis effect for the important traits (e.g. yield, earliness). New inbred lines are therefore screened for their heterotic potential by crossing them with a tester line belonging to a complementary heterotic group. The resulting F_1 hybrids are tested in extensive multi-environment field trials that absorb a fair amount of time, labour and budget. Therefore, despite the fact that experienced maize breeders tend to have some feeling on what combination of inbred lines is likely to demonstrate a commercially valuable heterosis effect, the development of a new hybrid variety generally remains a costly trial-and-error process.

It is clear that a hybrid breeding programme would benefit substantially from a reliable way to predict the phenotypic performance of an F_1 hybrid using only observable properties of its parents. Early prediction attempts use a phenotypic index that is obtained by measuring morphological traits on each of the parents, which is then used as a predictor for the yield of their F_1 maize hybrid (Anderson and Brown, 1952). The resulting R^2 values are however too low (16%) to be of any practical value. Other attempts try to exploit the observed correlation between the level of heterosis manifested by the hybrid

and the genetic diversity between its parents. Rao (1952) proposes to use the Mahalanobis D^2-statistic obtained from the measurements on one or more agronomic traits of the two parents in several environments, as an indirect measure of genetic diversity and as such, a predictor for heterosis. Again, the obtained prediction accuracy is of little practical value to the breeders.

From the early seventies to the late eighties, several papers were published concerning the use of isozymes for the estimation of genetic diversity and hybrid performance. Results for yield are generally disappointing as, for example, Smith and Smith (1989) make report of a prediction accuracy of $R^2 = 0.36$, using 32 isozyme loci on a panel of 100 maize hybrids. By contrast, a genetic diversity measure obtained by scoring 230 RFLP marker loci on the same panel results in an R^2 value of 0.87. This striking result is unfortunately only attainable for crosses between closely related inbred lines, as demonstrated by Melchinger et al. (1990). The correlation between the RFLP-based genetic distance measure and the phenotypic performance of crosses between unrelated lines (i.e. the standard approach in hybrid breeding programmes) is too weak for a reliable prediction of hybrid yield performance. This conclusion was reached simultaneously by Godshalk et al. (1990) and confirmed theoretically by Charcosset et al. (1991), Charcosset and Essioux (1994) and Bernardo (1992).

Bernardo (1994) makes predictions on single-cross hybrids having unrelated parents by means of Best Linear Prediction (BLP). The idea is to predict the phenotypic performance of untested hybrids by taking into account the observed phenotypic measurements of closely related hybrids. The latter are analysed by means of a linear mixed model where the different variance components (i.e. additive, dominance and residual variances) are estimated by means of Restricted Maximum Likelihood (REML). The covariance between the genotypic values of two hybrids is modelled according to a simplification of the covariance model described by Stuber and Cockerham (1966) by assuming linkage equilibrium between QTLs, statistical independence of identical alleles in the two involved heterotic groups and absence of epistasis and other higher order interactions between alleles. This covariance model requires estimates for the coefficient of coancestry (CoC) between all involved inbred lines within the same heterotic group. The required CoC estimates can be obtained from pedigree information but also by fingerprinting the inbred lines with co-dominant molecular markers as demonstrated earlier by Bernardo (1993).

The prediction accuracy of the presented BLP-based approach is promising for the important trait grain yield (i.e. maximum cross-validation-based $R^2 = 0.64$) and the RFLP-based covariance model seems to be slightly better than the pedigree-based model. In a series of subsequent papers, Bernardo (1995, 1996a,b,c) examines the robustness of the approach

by demonstrating applications to the unbalanced data generated by a commercial maize breeding programme and studying different traits, heterotic patterns, genetic model assumptions and misspecifications of the coefficient of coancestry. The general conclusion is that the BLP approach is robust, giving a relatively good prediction accuracy that is largely dependent on the heritability of the trait. Unfortunately, predictions on the much desired SCA component remain unreliable. Charcosset et al. (1998) use silage maize data to compare the capabilities of three different methods (i.e. genetic distance, principal component analysis and BLP) for improving the classic additive model by adding a predicted SCA component. For crosses between unrelated inbred lines, the improvements over the additive model are slim, but the BLP approach always results in the highest prediction accuracy.

The BLP approach, advocated by Bernardo, summarises the molecular marker scores of the hybrids in pairwise CoC estimates which are subsequently used to model the covariance between the hybrids. Other published prediction methods try to estimate the QTL effects, associated with the genotyped markers, more directly by some form of least squares estimation of the marker effects (Vuylsteke et al., 2000; Schrag et al., 2006, 2007). The problems related to the curse of dimensionality (i.e. simultaneous estimation of marker effects is not possible if the number of genotyped markers exceeds the number of genotypes) are circumvented by using only a subset of informative markers. This preliminary screening is based on the iterative or stepwise application of a parametric or non-parametric statistical test. Finding an appropriate significance threshold is however not straightforward as the familywise error rate and problems related to the multicollinearity of marker effects increase substantially when the number of markers to screen is large.

The advent of single nucleotide polymorphism (SNP) markers, in combination with high-throughput technologies, has opened the door to large-scale genotyping. The number of genotyped markers is rapidly increasing while the cost per entry is decreasing (Bernardo, 2008). This ongoing trend has motivated the development of the genomewide or genomic selection approaches as introduced by Meuwissen et al. (2001). The main idea is to skip the error-prone marker selection (i.e. screening) part and to use prediction methods that are able to fit all marker-based predictors without being susceptible to the inherent dimensionality problems of large molecular fingerprints. This is achieved by fitting the markers as random effects in a linear mixed model. Meuwissen et al. (2001) show, by means of a simulation study, that assuming heterogeneous variances for these marker effects in a Bayesian perspective allows to further improve the prediction accuracy. Bernardo and Yu (2007) demonstrate how this approach can be integrated in a maize breeding programme. Gianola et al. (2006) and Gianola and van Kaam (2008) explore the concept of repro-

ducing kernel Hilbert spaces regression (RKHS) to incorporate massive amounts of SNP scores in a linear mixed model setting. This regression approach is very closely related to the ε-insensitive support vector machine regression (ε-SVR), for which the theoretical foundations are discussed in the next chapter. The use of specific kernel functions allows to model non-additive, epistatic interactions between marker alleles, irrespective of the number of genotyped loci. González-Recio et al. (2008) compare this approach with three other marker-based prediction methods using a dataset on mortality rates of broiler chickens. Although the obtained prediction accuracies are very low for all methods, most likely as a consequence of the low heritability of this trait ($h^2 < 0.05$), the RKHS approach is shown to be superior to all examined competitors.

CHAPTER 3

ε-insensitive Support Vector Machine Regression

Support Vector Machines are a recent development in statistical learning theory by Vapnik (1995). The foundations of these techniques depend heavily on the theory of reproducing kernel Hilbert spaces for which a short introduction is provided. This introduction is by no means complete, but merely tries to provide some insight into the basic ideas. Readers expecting a more detailed treatment of the topic are referred to Aronszajn (1950) and Berg et al. (1984).

3.1 Reproducing kernel Hilbert spaces

A vector space $\mathbb{V}$ is a space that contains elements called 'vectors' over a field $\mathcal{X}$ that supports two kinds of operations: addition of vectors and multiplication by scalars drawn from the same field $\mathcal{X}$. These operators must obey several axioms including associative, commutative, inverse, identity and distributive laws. We typically think of vectors as finite dimensional arrays of elements over fields like $\mathbb{R}$ or $\mathbb{C}$ but the set of functions from $\mathbb{R} \to \mathbb{R}$, denoted as $\mathbb{R}^{\mathbb{R}}$, also represents a vector space. In this case, we can define addition and multiplication as $(fg)(x) = f(x)g(x)$ and $(af)(x) = af(x)$ for $f, g \in \mathbb{R}^{\mathbb{R}}$ and $a \in \mathbb{R}$. Another example of a vector space of functions is $\mathbb{R}^{\mathbb{R}^n}$, the set of functions that take an n-dimensional vector over the field $\mathbb{R}$ as an argument. These vector spaces of functions are generally referred to as function spaces.

A metric vector space $\mathbb{V}$ is a space that has a distance metric defined $d(\boldsymbol{x}, \boldsymbol{y})$ with $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{V}$, which allows to assess the proximity of these elements. A metric vector space $\mathbb{V}$ is said to be complete if every Cauchy sequence of vectors in $\mathbb{V}$ has a limit that is also in $\mathbb{V}$. This

basically means that the space $\mathbb{V}$ contains all elements that one would expect and that there are no missing vectors (i.e. no holes in the space). A Banach space is a complete vector space which is equipped with one or more vector norms. The norm of a vector $\boldsymbol{v}$, denoted as $\|\boldsymbol{v}\|$, allows to assess the size of this vector and has to satisfy the following properties:

$$\|\boldsymbol{v}\| \geq 0$$

$$\|\boldsymbol{v}\| = 0 \Leftrightarrow \boldsymbol{v} = \boldsymbol{0}$$

$$\|a\boldsymbol{u}\| = |a|\|\boldsymbol{u}\|$$

$$\|\boldsymbol{u} + \boldsymbol{v}\| \leq \|\boldsymbol{u}\| + \|\boldsymbol{v}\|$$

where $\boldsymbol{u}$, $\boldsymbol{v} \in \mathbb{V}$ and $a \in \mathcal{X}$. An example of such a norm defined for the complete vector space $\mathbb{R}^n$ is the p-norm where $p \geq 1$:

$$\|\boldsymbol{x}\|_p = \left(\sum_{i=1}^{n} |x_i|^p \right)^{\frac{1}{p}} .$$

If $p = 2$, this equation results in the well-known Euclidean norm. For complete function spaces that contain only continuous functions, a similar norm function can be defined as:

$$\|f\|_p = \left(\int_{-\infty}^{+\infty} |f^p(x)| dx \right)^{\frac{1}{p}} .$$

If we require this norm function to give a finite, positive size to all continuous functions f in the space, we can define such a space as

$$L_p = \{(f : \mathbb{R}^n \to \mathbb{R}) : \int_{-\infty}^{+\infty} |f^p(x)| dx < +\infty\} .$$

A Hilbert space $\mathcal{H}$ is a Banach space for which a dot-product operation is defined. If $\mathcal{H}$ is a vector space over the field $\mathcal{X}$, the result of this dot product is an element of $\mathcal{X}$. The dot product of $\boldsymbol{x}$ and $\boldsymbol{y}$ is denoted $\langle \boldsymbol{x}, \boldsymbol{y} \rangle$ and must satisfy the associative, commutative and distributive laws. $\mathbb{R}^n$ for example, is a Hilbert space for which the dot product for vectors $\boldsymbol{x}$ and $\boldsymbol{y} \in \mathbb{R}^n$ is defined as

$$\langle \boldsymbol{x}, \boldsymbol{y} \rangle = \sum_{i=1}^{n} x_i y_i .$$

In a similar way, we can define a dot product for functions f and $g \in \mathbb{R}^{\mathbb{R}^n}$ as:

$$\langle f, g \rangle = \int_{-\infty}^{+\infty} f(x)g(x)dx \,.$$

To explain the specifics of a reproducing kernel Hilbert space (RKHS), we first need to introduce the concept of functionals. Similar to an ordinary function $y = f(\boldsymbol{x})$ that is defined on the space $\mathbb{R}^n$ and takes values in $\mathbb{R}$, a functional T is a function on a space of functions $\mathbb{V}$ that determines uniquely a number in $\mathbb{R}$ for each member in $\mathbb{V}$. For example, if $\mathbb{V} = L_2(\mathcal{X})$ (i.e. the L_2 space of functions that take an argument from the field $\mathcal{X}$), a functional T over $\mathbb{V}$ could take the form

$$T(f) = \int_{\mathcal{X}} f(x)dx \,.$$

A functional is linear if it satisfies

$$T(\alpha f + \beta g) = \alpha T(f) + \beta T(g) \,,$$

for all real numbers α and β and all members f and g in $\mathbb{V}$. A functional F is bounded if there exists a number $M > 0 \in \mathbb{R}$ such that for all $f \in \mathbb{V}$:

$$\|F(f)\| \leq M\|f\| \,.$$

The Riesz representation theorem states that if F is a bounded linear functional on a Hilbert space $\mathcal{H}$, then there is always a unique vector (function) $g \in \mathcal{H}$ for which

$$F(f) = \langle f, g \rangle \,.$$

A Dirac evaluation functional $F_{\boldsymbol{x}}$ is a bounded linear functional defined as

$$F_{\boldsymbol{x}}(f) = f(\boldsymbol{x}), \quad \forall f \in \mathcal{H} \,,$$

which basically evaluates the argument function f at the point $\boldsymbol{x} \in \mathcal{X}$ (where generally $\mathcal{X} \subseteq \mathbb{R}^n$). A RKHS is defined as a Hilbert space for which all Dirac evaluation functionals $F_{\boldsymbol{x}}$ are bounded. According to the Riesz representation theorem this implies that

$$F_{\boldsymbol{x}}(f) = f(\boldsymbol{x}) = \langle f, k_{\boldsymbol{x}} \rangle \,,$$

where $k_{\boldsymbol{x}} \in \mathcal{H}$ is the unique representer function for the Dirac evaluation functional $F_{\boldsymbol{x}}$ at a point $\boldsymbol{x} \in \mathcal{X}$. As $k_{\boldsymbol{x}}$ is a function itself, we can evaluate it at a different point $\boldsymbol{y} \in \mathcal{X}$ as

$$F_{\boldsymbol{y}}(k_{\boldsymbol{x}}) = k_{\boldsymbol{x}}(\boldsymbol{y}) = \langle k_{\boldsymbol{y}}, k_{\boldsymbol{x}} \rangle \,.$$

Using this, we can define the reproducing kernel function K for $\mathcal{H}$ as $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ or $K(\boldsymbol{x}, \boldsymbol{y}) = \langle k_{\boldsymbol{x}}, k_{\boldsymbol{y}} \rangle$ where $\boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}$. The reproducing property of such a kernel function relates to the fact that the dot product $\langle f(.), K(\boldsymbol{x}, .) \rangle = \langle f, k_{\boldsymbol{x}} \rangle = f(\boldsymbol{x})$ reproduces the function f evaluated at element $\boldsymbol{x}$. It can be shown that this reproducing kernel function is unique for each RKHS. Mercer's theorem states the properties of a symmetric function $K(\boldsymbol{x}, \boldsymbol{y})$ to be a reproducing kernel function with an associated RKHS.

Theorem 3.1, **Mercer's Theorem** (Mercer, 1909; Cristianini and Shawe-Taylor, 2000): Let $\mathcal{X}$ be a compact subset of $\mathbb{R}^n$. Suppose $K(\cdot, \cdot)$ is a continuous symmetric function such that the integral operator $T_K : L_2(\mathcal{X}) \to L_2(\mathcal{X})$,

$$(T_K f)(\cdot) = \int_{\mathcal{X}} K(\cdot, \boldsymbol{x}) f(\boldsymbol{x}) d\boldsymbol{x} \,,$$

is positive, that is

$$\iint_{\mathcal{X} \times \mathcal{X}} K(\boldsymbol{x}, \boldsymbol{y}) f(\boldsymbol{x}) f(\boldsymbol{y}) d\boldsymbol{x} d\boldsymbol{y} \geq 0 \,,$$

for all $f \in L_2(\mathcal{X})$. Then we can expand $K(\boldsymbol{x}, \boldsymbol{y})$ in a uniformly convergent series on $\mathcal{X} \times \mathcal{X}$ in terms of T_K's orthonormal eigenfunctions $\omega_j \in L_2(\mathcal{X})$, normalised in such a way that $\|\omega_j\|_{L_2} = 1$, and positive associated eigenvalues $\lambda_j \geq 0$,

$$K(\boldsymbol{x}, \boldsymbol{y}) = \sum_{j=1}^{d_{\mathcal{H}}} \lambda_j \omega_j(\boldsymbol{x}) \omega_j(\boldsymbol{y}) \,,$$

where $d_{\mathcal{H}}$ is the dimension of the Hilbert space, either $d_{\mathcal{H}} \in \mathbb{N}$ or $d_{\mathcal{H}} = \infty$.

Kernel functions that obey this positivity condition are called Mercer kernels and it can be shown that for these kernels, there always exists a RKHS $\mathcal{H}$ of functions defined over $\mathcal{X}$ for which K is the reproducing kernel. The converse is also true, meaning that for any Hilbert space of functions in which the Dirac evaluation functionals are bounded and linear (i.e. any RKHS), there exists an associated reproducing kernel function which is also a Mercer kernel (Wahba, 1990; Cristianini and Shawe-Taylor, 2000). The positivity condition of the integral operator T_K is a generalisation of the positive semi-definite (PSD) property of the kernel matrix obtained for any finite subset of $\mathcal{X}$ of size m as

$$\sum_{i=1}^{m} \sum_{j=1}^{m} a_i a_j K(x_i, x_j) \geq 0, \quad \forall a_i, a_j \in \mathbb{R} \,,$$

Mercer's theorem states that each symmetric, PSD kernel function K has an associated RKHS $\mathcal{H}$ for which K is the unique, reproducing kernel function. In addition to $\mathcal{H}$, we

can define a mapping ϕ_m between each element of $\mathcal{X}$ and its associated vector in a $d_{\mathcal{H}}$-dimensional vector space $\mathcal{F}$ called the feature space as

$$\phi_m : \mathcal{X} \subseteq \mathbb{R}^n \to \mathcal{F} : \boldsymbol{x} \to [\sqrt{\lambda_1}\omega_1(\boldsymbol{x}) \dots \sqrt{\lambda_{d_{\mathcal{H}}}}\omega_{d_{\mathcal{H}}}(\boldsymbol{x})] .$$

The dot product of two of these vectors in the feature space can be represented as

$$\begin{aligned} \langle \phi_m(\boldsymbol{x}), \phi_m(\boldsymbol{y}) \rangle &= \sum_{i=1}^{d_{\mathcal{H}}} \lambda_i \omega_i(\boldsymbol{x}) \omega_i(\boldsymbol{y}) \\ &= K(\boldsymbol{x}, \boldsymbol{y}) \end{aligned}$$

This result is in fact the main conclusion of our introduction to the concept of a RKHS. Any symmetric, PSD kernel function K when evaluated for two vectors $\boldsymbol{x}$ and $\boldsymbol{y}$ basically evaluates the dot product between the mapping of these two vectors in some (preferably high dimensional) feature space. It is however important to realise that the RKHS induced by a particular PSD kernel function K is only unique up to isometric isomorphism which implies that the mapping ϕ is also not unique. Therefore, the mapping ϕ_m carries the subscript m to indicate that this mapping is a direct consequence of Mercer's theorem. Other RKHSs and mappings can be described for the same kernel function K, but there always exists an isometric isomorphism between their induced feature spaces (Schölkopf and Smola, 2001).

3.2 Structural Risk Minimisation

The ε-insensitive Support Vector Machines regression (ε-SVR) technique was built upon the statistical learning framework developed by Vapnik (1995, 1998). This framework assumes that there are two sets of variables $\boldsymbol{x} \in \mathcal{X} \subseteq \mathbb{R}^n$ and $y \in \mathcal{Y} \subseteq \mathbb{R}$ that are related by a probability distribution $p(\boldsymbol{x}, y)$ over the set $\mathcal{X} \times \mathcal{Y}$. This probability distribution is unknown but we assume to have a random sample of this distribution $\mathcal{D}_m = ((\boldsymbol{x}_i, y_i))_{i=1}^m \subseteq \mathcal{X} \times \mathcal{Y}$ for which it is assumed that the measurement error on the vectors $\boldsymbol{x}_i$ is negligible compared to the accuracy of the y_i measurements. We want to find a function $f : \mathcal{X} \to \mathcal{Y}$ using the data set $\mathcal{D}_m$, which we can use to predict the value y for any given $\boldsymbol{x} \in \mathcal{X}$.
To find this function f, we need to define a set of candidates which are grouped in what is called the hypothesis space $\mathcal{F}$. We also need a criterion which allows to select the best candidate function from this hypothesis space $\mathcal{F}$. In statistical learning theory, this criterion is called a risk functional, which measures the average error associated with an estimator f. If the function $V(y, f(\boldsymbol{x}))$ is a loss function measuring the error we make

when we predict y by $f(\boldsymbol{x})$, then the theoretical or expected risk functional can be defined as

$$I[f] = \iint_{\mathcal{X}\times\mathcal{Y}} V(y, f(\boldsymbol{x}))p(\boldsymbol{x}, y)d\boldsymbol{x}dy\,.$$

Unfortunately it is not possible to find the function f that minimises this theoretical risk functional (often called the target function f_0) because the probability distribution $p(\boldsymbol{x}, y)$ is unknown. The empirical risk minimisation method therefore uses the data set $\mathcal{D}_m$ to build a stochastic approximation of the expected risk, which is usually called the empirical risk, and is defined as:

$$I_{emp}[f, \mathcal{D}_m] = \frac{1}{m}\sum_{i=1}^{m} V(y_i, f(\boldsymbol{x}_i))\,. \tag{3.1}$$

Straightforward minimisation of the empirical risk is generally problematic as there are an infinite number of solutions and it can lead to overfitting, meaning that although the minimum of the empirical risk can be very close to zero, the theoretical risk, which is what we are really interested in, can be very large (Evgeniou et al., 2002). One can however define a probabilistic bound on the distance between the empirical and theoretical risk of a function f. This bound involves the number of examples m and the capacity h of the function space. This capacity h is a measure for the complexity of the hypothesis space $\mathcal{F}$, or in other words, a measure for the flexibility of the functions in $\mathcal{F}$. Several appropriate capacity measures have been proposed in literature but the most general one is the Vapnik-Chervonenkis (VC) dimension (Vapnik and Chervonenkis, 1971). This VC-dimension is defined as the maximum number of training examples that can be separated by the members of the hypothesis space, for each possible labelling of the points. For example, the set of linear functions in $\mathbb{R}^n$ has a VC-dimension $h = n + 1$ since it is not possible to separate more than $n + 1$ points by a linear hyperplane in $\mathbb{R}^n$ for each possible labelling. The higher the VC-dimension, the easier it becomes to reduce the empirical risk, at the cost of an increased probability that the minimiser overfits the training data. The probabilistic bound on the distance between the empirical and the theoretical risk of a function f obeys with a probability η (Evgeniou et al., 2002):

$$I[f] < I_{emp}[f, \mathcal{D}_m] + \varphi\left(\sqrt{\frac{h}{m}}, \eta\right)\,, \tag{3.2}$$

where φ is an increasing function of $\frac{h}{m}$ and η. From this equation it should be clear that a minimisation of both empirical risk and capacity of the hypothesis space are needed to minimise theoretic risk. These are conflicting objectives as a higher capacity h will make it easier to fit the training data (i.e. to minimise the empirical risk I_{emp}), while function

sets with lower capacity tend to generalise more, which results in an increased empirical risk.

The idea of structural risk minimisation is to define a nested sequence of hypothesis spaces $\mathcal{H}_1 \subset \mathcal{H}_2 \subset \ldots \subset \mathcal{H}_p$, where each hypothesis space $\mathcal{H}_i$ has a finite capacity h_i that is larger than that of all previous sets, that is $h_1 \leq h_2 \leq \ldots \leq h_p$. One needs to identify the function f that minimises the empirical risk in the hypothesis space $\mathcal{H}_i$ that minimises the right-hand side of Eq. (3.2).

In a RKHS, a nested sequence of functions can be constructed by bounding the norm of the functions. This can be achieved by defining a set of constants $a_1 \leq a_2 \leq \ldots \leq a_p$ which allows to define a nested series of p function spaces of the form (Evgeniou et al., 1999)

$$\mathcal{H}_i = \{f \in RKHS : \|f\| \leq a_i\}\,.$$

Unfortunately, solving the constrained optimisation problem for each value a_i is generally unfeasible as in theory, a_i can take an infinite number of monotonically increasing values. To bypass this issue, the problem is reformulated as finding the function f that minimises (Tikhonov and Arsenin, 1977; Evgeniou et al., 2002)

$$\min_{f \in \mathcal{H}} \frac{1}{m} \sum_{i=1}^{m} V(y_i, f(\boldsymbol{x}_i)) + \lambda \|f\|^2\,. \tag{3.3}$$

This equation allows for a smooth trade-off between the empirical risk of the training data $\mathcal{D}_m$ and the size of the RKHS norm of the set of candidate functions, an indirect measure for their capacity h. The regularisation parameter λ is basically a penalty for functions with a high capacity: the larger λ, the smaller the allowed RKHS norm of the solution function. The optimal value for λ is usually determined by some form of cross-validation on the training data.

There are many possible choices for the loss function $V(y, f(\boldsymbol{x}))$ but the most common variants for regression purposes are:

the L_1 norm loss function: $V(y, f(\boldsymbol{x})) = |y - f(\boldsymbol{x})|$

the L_2 norm or squared loss function: $V(y, f(\boldsymbol{x})) = (y - f(\boldsymbol{x}))^2$

the ε-insensitive loss function: $V(y, f(\boldsymbol{x})) = \max(0, |y - f(\boldsymbol{x})| - \varepsilon)$

the quadratic ε-insensitive loss function: $V(y, f(\boldsymbol{x})) = \max(0, (|y - f(\boldsymbol{x})| - \varepsilon)^2)$

3.3 Linear ε-SVR

ε-SVR is intrinsically a non-parametric, non-linear regression technique. We initially demonstrate the approach by restricting the set of candidate functions to linear functions of the form

$$f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle + b \,,$$

where $\boldsymbol{w} \in \mathbb{R}^n$ and $b \in \mathbb{R}$. For the moment, we will assume that there actually exists such a linear function which, according to the ε-insensitive loss function, results in a perfect fit of the training data $\mathcal{D}_m$. This means that all training examples lie within a distance of ε of the hypersurface created by f which translates into

$$\begin{aligned} y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b &\leq \varepsilon \\ \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - y_i &\leq \varepsilon \,, \end{aligned}$$

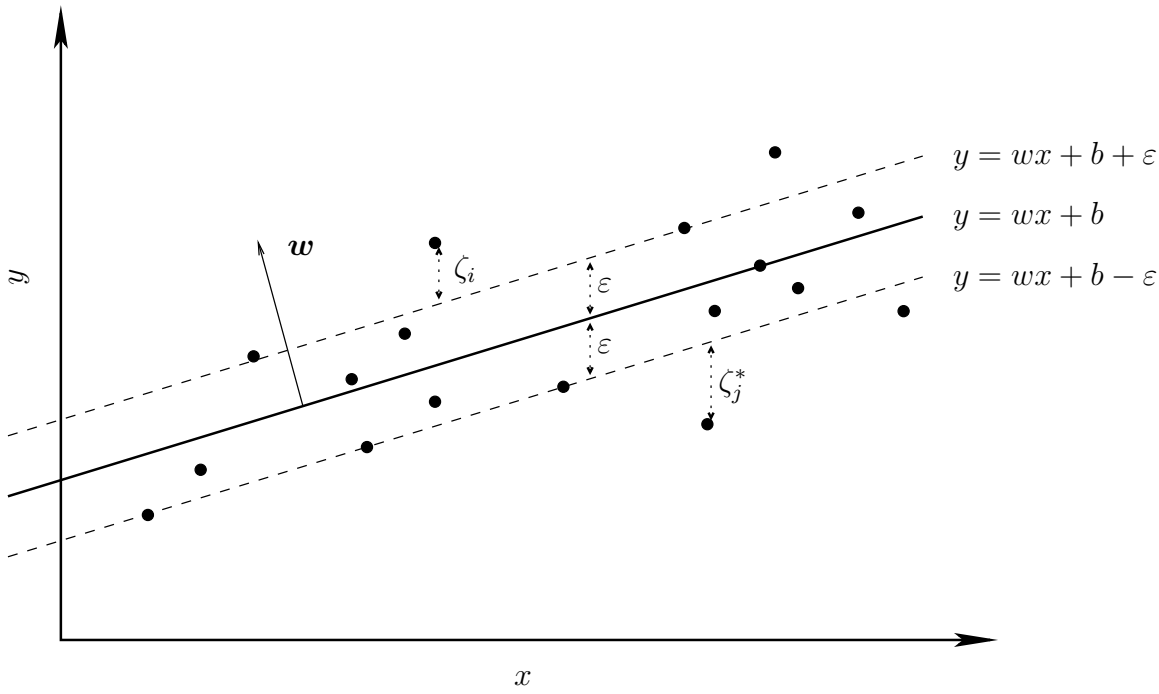
for all $(\boldsymbol{x}_i, y_i) \in \mathcal{D}_m$. Figure 3.1 gives a graphical representation of such a linear function f in $\mathbb{R}^2$. We can see that not all points lie within the grey insensitivity tube of width $\frac{2\varepsilon}{\|\boldsymbol{w}\|}$ which means that the depicted function f does not provide a perfect fit to the training data according to the ε-insensitive loss function. To accommodate these non-fitting points, we introduce slack variables ζ and ζ^* which allow the function f to respectively underestimate or overestimate the actual value y of a training example $(\boldsymbol{x}, y)$. Using these slack variables we can reformulate the minimisation problem of Eq. (3.3) using the ε-insensitive loss function as (Vapnik, 1995)

$$\min_{\boldsymbol{w},b} \frac{1}{2}\|\boldsymbol{w}\|^2 + C \sum_{i=1}^{m} (\zeta_i + \zeta_i^*) \,, \tag{3.4}$$

$$\text{subject to} \begin{cases} y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b & \leq \varepsilon + \zeta_i \\ \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - y_i & \leq \varepsilon + \zeta_i^* \\ \zeta_i, \zeta_i^* & \geq 0 \end{cases}$$

The variable C plays the role of λ in Eq. (3.3) and allows to make a trade-off between the capacity (or flatness) of the function and the prediction errors it makes on the training data in $\mathcal{D}_m$. Only half of the norm of vector $\boldsymbol{w}$ is used as this allows for a cleaner formulation of the problem after differentiation with respect to $\boldsymbol{w}$. Eq. (3.4) is called the primal formulation of the problem.

Figure 3.1: Example of a linear function $y = f(x) = wx+b$ in combination with the insensitivity tube of width $\frac{2\varepsilon}{\|\boldsymbol{w}\|}$ depicted by the dashed lines. The points scattered around the regression line are the training examples $(x_i, y_i) \in \mathcal{D}_m \subset \mathbb{R}^2$. Points within the insensitivity tube are considered to have a zero error when predicted by the linear function f according to the ε-insensitive loss function.



The inequality constraints are included into the minimisation problem by means of Lagrange multipliers which results in what is called the Lagrangian of the problem:

$$\begin{aligned} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\eta}) =& \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{m}(\zeta_i + \zeta_i^*) - \sum_{i=1}^{m}(\eta_i\zeta_i + \eta_i^*\zeta_i^*) \\ & - \sum_{i=1}^{m}\alpha_i(\varepsilon + \zeta_i - y_i + \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b) \\ & - \sum_{i=1}^{m}\alpha_i^*(\varepsilon + \zeta_i^* + y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b) \end{aligned}$$

The Lagrange multipliers $\eta_i^{(*)}$ and $\alpha_i^{(*)}$ in L are called the dual variables. As all constraints are linear, we can invoke the Karush-Kuhn-Tucker (KKT) conditions (Nocedal and Wright, 1999), an important result from the field of constrained optimisation, which state that at the optimal values for $\boldsymbol{w}$ and b, the following conditions are met:

$$\partial_{\boldsymbol{w}} L = \boldsymbol{w} - \sum_{i=1}^{m}(\alpha_i - \alpha_i^*)\boldsymbol{x}_i = 0 \tag{3.5}$$

$$\partial_b L = \sum_{i=1}^{m}(\alpha_i - \alpha_i^*) = 0 \tag{3.6}$$

$$\partial_{\zeta_i^{(*)}} L = C - \alpha_i^{(*)} - \eta_i^{(*)} = 0 \tag{3.7}$$

$$\zeta_i^{(*)} \geq 0\,, \quad \forall i = 1, \ldots, m \tag{3.8}$$

$$y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b - \varepsilon - \zeta_i \leq 0\,, \quad \forall i = 1, \ldots, m \tag{3.9}$$

$$\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b - y_i - \varepsilon - \zeta_i^* \leq 0\,, \quad \forall i = 1, \ldots, m \tag{3.10}$$

$$\eta_i^{(*)} \geq 0\,, \quad \forall i = 1, \ldots, m \tag{3.11}$$

$$\alpha_i^{(*)} \geq 0\,, \quad \forall i = 1, \ldots, m \tag{3.12}$$

$$\eta_i^{(*)}\zeta_i^{(*)} = 0\,, \quad \forall i = 1, \ldots, m \tag{3.13}$$

$$\alpha_i(y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b - \varepsilon - \zeta_i) = 0\,, \quad \forall i = 1, \ldots, m \tag{3.14}$$

$$\alpha_i^*(\langle \boldsymbol{w}, \boldsymbol{x}_i\rangle + b - y_i - \varepsilon - \zeta_i^*) = 0\,, \quad \forall i = 1, \ldots, m \tag{3.15}$$

The first three conditions state that each partial derivative of the Lagrangian, evaluated at the optimal tuple $(\boldsymbol{w}, b)$ is zero. Eqs. (3.9)-(3.11) indicate that the solution is feasible (i.e. all constraints are met) and Eqs. (3.12)-(3.14) indicate that each of the dual variables is either positive or zero. The last three conditions state that each of the dual variables is necessarily zero if their equivalent constraint is not active. This means for example that α_i is zero if $y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i\rangle - b - \varepsilon - \zeta_i$ is not, a condition only met by points which are located

on or above the hypersurface and within the insensitivity tube. If on the other hand the point is located on the boundary of the insensitivity tube, $y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - b - \varepsilon - \zeta_i$ becomes zero which indicates that the dual variable α_i can have a non-zero value. As the objective function of Eq. (3.4) is convex and the inequality constraints are linear (i.e. they give rise to a convex feasible region), the KKT conditions can be shown to be both necessary and sufficient conditions for a tuple $(\boldsymbol{w}, b)$ to be the optimal solution to the minimisation problem.

As a result of the convexity of both the objective function and its constraints, we can also make use of the strong duality theorem which states that minimising the primal objective function of Eq. (3.4) is equivalent to maximising the dual optimisation problem with respect to the dual variables:

$$\max_{\boldsymbol{\alpha},\boldsymbol{\eta}} \min_{\boldsymbol{w},b} L(\boldsymbol{w}, b, \boldsymbol{\alpha}, \boldsymbol{\eta})$$
$$\text{subject to} \begin{cases} \alpha_i^{(*)} \geq 0\,, & \forall i = 1 \ldots m \\ \eta_i^{(*)} \geq 0\,, & \forall i = 1 \ldots m \end{cases}$$

The minimum of the Lagrangian is obtained at the vanishing partial derivatives with respect to the unknowns $\boldsymbol{w}$, b and $\zeta_i^{(*)}$ (i.e. Eqs. (3.6)-(3.8)) which provides the expressions $\boldsymbol{w} = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \boldsymbol{x}_i$ and $\eta_i^{(*)} = C - \alpha_i^{(*)}$. If we substitute these expressions back into the Lagrangian we obtain the dual optimisation problem:

$$\begin{aligned} \max_{\boldsymbol{\alpha}} & -\frac{1}{2} \sum_{i=1}^{m} \sum_{j=1}^{m} (\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*) \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle \\ & - \varepsilon \sum_{i=1}^{m} (\alpha_i + \alpha_i^*) + \sum_{i=1}^{m} y_i (\alpha_i - \alpha_i^*)\,, \end{aligned} \tag{3.16}$$

subject to the constraints $\sum_{i=1}^{m} (\alpha_i - \alpha_i^*) = 0$ and $\alpha_i^{(*)} \in [0, C]$. Note that Eq. (3.6) allows us to rewrite f as:

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \langle \boldsymbol{x}_i, \boldsymbol{x} \rangle + b\,. \tag{3.17}$$

Eq. (3.17) shows that f is specified as a linear combination of all training examples $\boldsymbol{x}_i$ expressed as dot products. We also deduced from the KKT conditions that the coefficients α_i or α_i^* can only be non-zero if $(y_i - f(\boldsymbol{x}_i)) \geq \varepsilon$ or $(f(\boldsymbol{x}_i) - y_i) \geq \varepsilon$ respectively. This means that all samples inside the insensitivity tube are not used in the formulation of f. All other training examples with nonvanishing coefficients $\alpha_i^{(*)}$ are called the support vectors, hence the name Support Vector Machines. This approach allows for a sparse representation of function f, which makes calculating $f(\boldsymbol{x})$ very efficient.

Now imagine we have found the values for α_i and α_i^* that maximise Eq. (3.16). If we want to represent $f(\boldsymbol{x})$ according to Eq. (3.17), we still need an estimate for variable b. There are several approaches for obtaining the optimal value for b but these generally depend on the optimisation routine that is used for maximising Eq. (3.16) (Keerthi et al., 2001). A more theoretical approach is based on the the KKT conditions. If $\alpha_i > 0$ (i.e. point i is on the boundary of the insensitivity tube or above), α_i^* necessarily equals 0 as otherwise the the last KKT condition, $\alpha_i^*(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle + b - y_i - \varepsilon - \zeta_i^*) = 0$ will not be fulfilled. A similar reasoning for $\alpha_i^* > 0$ allows to conclude that $\alpha_i \alpha_i^* = 0$ for all points $i \in \mathcal{D}_m$. Now imagine that $\alpha_i < C$, then we know from Eq. (3.8) that $\eta_i > 0$ which means that according to Eq. (3.14), $\zeta_i = 0$ which in turn implies (using Eq. (3.10)) that $b \geq y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i, - \rangle \varepsilon$. If on the other hand $\alpha_i > 0$, we know that $\alpha_i^* = 0$ which results in $b \leq y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i, + \rangle \varepsilon$. A similar reasoning for α_i^* allows for the statement

$$\max_{i \in \mathcal{D}_m} \{ y_i - \langle \boldsymbol{w}, \boldsymbol{x}_i \rangle - \varepsilon \mid \alpha_i < C \text{ or } \alpha_i^* > 0 \} \leq b \leq \min_{j \in \mathcal{D}_m} \{ y_j - \langle \boldsymbol{w}, \boldsymbol{x}_j \rangle + \varepsilon \mid \alpha_j^* < C \text{ or } \alpha_j > 0 \} . \tag{3.18}$$

If there exists an α_i for which $0 < \alpha_i < C$ or equivalently an α_i^* for which $0 < \alpha_i^* < C$, the inequalities in Eq. (3.18) become equalities and as such, allow for the estimation of b.

3.4 Extension to non-linear models

If we preprocess the training examples $\boldsymbol{x}_i$ by a map $\phi : \mathcal{X} \to \mathcal{F}$ into a higher dimensional space, i.e. the feature space $\mathcal{F}$, and solve the linear regression there, we can state Eq. (3.17) as

$$f(\boldsymbol{x}) = \sum_{i=1}^{m} (\alpha_i - \alpha_i^*) \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle + b . \tag{3.19}$$

Depending on the map ϕ this approach effectively allows us to create non-linear functions f. When predicting y for an unknown example $\boldsymbol{x}$ using the in feature space learned linear function f, Eq. (3.19) obliges us to apply the mapping ϕ to this new case as well as to all training examples and subsequently make the dot product between them. This approach is often not computationally feasible, so we use instead a symmetric kernel function $K(\boldsymbol{x}_i, \boldsymbol{x}) = \langle \phi(\boldsymbol{x}_i), \phi(\boldsymbol{x}) \rangle$ that gives us directly the dot product in some RKHS. This shortcut allows us to reformulate Eq. (3.19) as

$$f(\boldsymbol{x}) = \sum_{i=1}^{l} (\alpha_i - \alpha_i^*) K(\boldsymbol{x}_i, \boldsymbol{x}) + b . \tag{3.20}$$

Note how also the dual optimisation function of Eq. (3.16) is based on dot products between training vectors $\boldsymbol{x}_i$, which makes it straightforward to plug in a well-chosen kernel function in the formulation. This approach is commonly called the 'kernel trick'.
Some commonly used kernel functions are:

the polynomial kernel with degree d: $K(\boldsymbol{x}, \boldsymbol{z}) = (\boldsymbol{x}'\boldsymbol{z} + 1)^d$

the radial basis or Gaussian kernel with kernel width γ: $K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma\|\boldsymbol{x} - \boldsymbol{z}\|^2)$

the sigmoid kernel with parameters κ and θ: $K(\boldsymbol{x}, \boldsymbol{z}) = \tanh(\kappa\boldsymbol{x}'\boldsymbol{y} + \theta)$

most kernel functions require the knowledge of one or several kernel parameters. Furthermore, in the derivation of Eqs. (3.16) and (3.17) it was also assumed that C and ε are known parameters. In rare cases, good values for all these parameters can be copied from previous studies on similar training data but in general, some form of multi-dimensional grid-search in combination with a cross-validation-based target score function is required.

3.5 Sequential minimal optimisation

So far, it has been assumed that we can easily find the dual variables $\alpha_i^{(*)}$ that maximise Eq. (3.17). This task is unfortunately not that straightforward, as we are dealing with a constrained, quadratic programming problem (QP) with generally, a very high dimensionality. One rather simple approach consists of iteratively selecting arbitrary subsets of the data, denoted as the working set, and subsequently solving the problem for these examples (Boser et al., 1992). Only supporting vectors (i.e. training points where $\alpha_i^{(*)} > 0$) are added to the next chunk, and the process is repeated until all support vectors are found. Unfortunately, this chunking approach still cannot handle large-scale training problems if the number of support vectors is large. Osuna et al. (1997) prove that a decomposition of such a problem in smaller, more manageable pieces, will reduce the overall objective function as long as at least one example that violates the KKT conditions is added to the working set at each iteration. It should, however, be clear that a each decomposed piece of the problem remains a QP which needs to be solved by for example stochastic gradient descent, Newton's method, conjugate gradient descent or a primal-dual interior point method.
Platt (1998) takes decomposition to the extreme by only allowing a working set of size two. This problem can be solved analytically, avoiding a numerical QP solver altogether. He calls this approach sequential minimal optimisation (SMO) and as long as the two selected

variables at each iteration violate the KKT conditions, the algorithm is guaranteed to converge to the optimal solution, in accordance with Osuna's theorem. In practice however, the algorithm suffers from slow convergence if the working set is chosen at random. Keerthi et al. (2001) provide an improvement to Platt's algorithm by specifically using the 'maximal violating pair' at each iteration. The identification of this maximal violating pair is based on Eq. (3.18), a direct result of the KKT conditions. This approach is shown to result in a significant improvement in the speed of convergence. Fan et al. (2005) show that the maximal violating pair is related to the first order approximation of Eq. (3.16) and show how the working set selection algorithm can be improved by means of a second order approximation of the optimisation function. This working set selection approach is integrated in the software package LIBSVM (Chang and Lin, 2006) which allows for both classification and regression using the support vector machines framework by an elegant reformulation of Eq. (3.16).

3.6 Conclusions

In this chapter, an attempt was made to introduce the theoretical foundations of ε-insensitive support vector machines regression. However, it needs to be emphasised that this overview is by no means claimed to be comprehensive and that ε-SVR itself, only represents a tiny speckle in the world of techniques that are based on the kernel trick and/or support vector machines. Most textbooks introduce these topics in the chronological sequence of their development, namely support vector machines as a non-parametric classification technique for which subsequently, numerous adaptions and optimisations have been published, including the presented ε-SVR. Other interesting developments include ν-support vector classification and regression (ν-SVM and ν-SVR) in which the number of support vectors (i.e. the sparsity) and training error are being controlled (Schölkopf et al., 2000) and least squares support vector machines (LS-SVM), in which the standard classification and regression problems are reformulated such that the optimisation problems become linear and therefore easier to solve (Suykens and Vandewalle, 1999).
Notwithstanding the plethora of elegant yet powerful kernel-based techniques, it was opted to focus this chapter directly on ε-SVR, as this is the only kernel-based technique that is explored in this dissertation. Furthermore, all presented results on ε-SVR-based hybrid prediction are obtained by using the linear, ε-insensitive loss function, while numerous other loss functions could have been envisioned. In a similar spirit of downsizing the option list, the performance of only a limited set of well-known kernel functions was explored.

These observations allow to conclude that the presented results and conclusions on hybrid prediction, should be considered as the low-hanging fruit of the new and continuously expanding world of kernel-based methods.

CHAPTER 4

Data description

This dissertation is the result of a close collaboration between University College Ghent, Ghent University, the Institute for Agricultural and Fisheries Research (ILVO) and the private breeding company RAGT R2n. The phenotypic data that was used in different parts of the presented research was generated as part of the grain maize breeding programme of RAGT R2n. The molecular marker data on the other hand was partially provided by RAGT R2n (SSR) and partially by the Plant Sciences Unit of ILVO (AFLP). Due to confidentiality agreements, this dissertation contains no details about the structure of the grain maize breeding pool of RAGT R2n nor does it contain technical aspects of their field and molecular experiments.

4.1 Phenotypic data

The major part of the inbred lines in the RAGT grain maize breeding pool has a corn belt dent background. The pool is roughly divided in several big heterotic groups, named according to historically important corn belt inbred lines or OPVs such as Iowa Stiff Stalk Synthetic (ISSS), Iodent Reid and Lancaster Sure Crop. Each of these heterotic groups is composed of several subheterotic groups, generally having a long history of reciprocal recurrent selection in an attempt to optimise the heterotic response for grain yield.
Each new combination of inbred lines is tested in one or several multi-environment trials (MET). To implement these METs, RAGT R2n has several experimental stations in France and other European countries. All phenotypic measurements that were recorded as part of the RAGT grain maize breeding programme from 1985 until 2005 were made available to this research. These measurements are grouped in 5197 METs, evaluating the phenotypic performance of on average 28.7 hybrids in on average 3.4 distinct locations. A MET always

starts and ends within a single growing season. We use the word trial when referring to a field experiment that measures the phenotypic performance of a set of hybrids at a particular location as part of a MET. Trials are therefore nested within METs and as such, also nested within growing seasons. A single location (i.e. experimental station) in a particular growing season generally houses trials belonging to different METs, but a single MET can also have more than one trial at the same location, altough this generally implies different experimental settings with varying sowing and harvesting dates or fertiliser and irrigation treatments.

The experimental design and number of replications varies between METs. In 61.6% of the trials, there is only a single replication while the experimental design in the remaining trials is always resolvable, allowing for a reduction to randomised complete block designs. A plot is the smallest experimental unit and allows to measure the phenotypic response of a single hybrid in a particular trial. For each plot, several phenotypic traits can be recorded including yield (in quintaux $\text{ha}^{-1} = 100$ kg ha^{-1}), grain moisture content (%), the number of days until female flowering, stalk and root lodging and ordinal scores on various pests and fungi including *Ostrinia nubilalis* (European corn borer), *Ustilago maydis* (corn smut) and *Helminthosporium spp.* Not all plots have a recording for each of these traits but the data for yield and moisture content are complete. The measurements for the trait days until flowering are generally not replicated within the same trial.

The budget for molecular fingerprinting was limited and therefore it was decided to include approximately 200 inbred lines in the presented study. Selecting these inbred lines from the huge set of candidates turned out to be a non-trivial task. A single inbred line is, on average, parent of 4.3 distinct hybrids but a lot of inbred lines only have a single offspring while others (i.e. the tester lines) are parent of thousands of hybrids. Moreover, a single hybrid is tested on average in 28.7 trials but many hybrids are only tested once, while others have served as checked varieties and are therefore present in numerous trials. The problem of identifying the optimal set of inbred lines with respect to the available unbalanced phenotypic data of their hybrids is tackled using a graph-based procedure which is discussed in Chapter 5. Unfortunately, this procedure was not yet finished at the early stages of the presented research and therefore the selection of 200 inbred lines was guided by a set of reasonable but nevertheless rather ad hoc conditions. The underlying idea is to maximise the number of hybrids for which both genotypic (i.e. molecular marker) and phenotypic information is available (i.e. the training set), within the constraints of the fixed genotyping budget. As all candidate inbred lines are almost completely homozygous, we can nearly always deduce the entire molecular fingerprint of a hybrid from the marker scores of its parents. This means that we need to select the set of 200 inbred lines which

has created the largest number of single-cross hybrids amongst themselves. On the other hand, hybrids which have only a small number of phenotypic records should not be taken into account as it will not be possible to obtain a reliable estimate of genotypic performance for these hybrids. These observations resulted in the following five conditions which are met by the selection of inbred lines used in the presented research:

The set of candidate inbred lines is limited to one specific combination of two (sub)heterotic groups denoted hereafter as ISSS and Iodent respectively. This specific heterotic combination has resulted in the largest number of hybrids for which phenotypic data is available and is therefore more likely to result in a training set of maximum size.

The hybrids under consideration should be tested in METs that were performed in or after the year 1998. Choosing the most recent field trials automatically results in selecting recently developed inbred lines which are already routinely being fingerprinted using SSR and SNP markers. Altough we need a higher marker density than what is provided by the routine fingerprinting procedure, the already available SSR marker scores save time and money. Moreover, older inbred lines are usually no longer available for fingerprinting. A second argument for this condition can be found in the observed connectivity structure of the phenotypic data. Trials which have no varieties in common could be disconnected and this is more likely for trials that are separated far in time. As will be explained in detail in Chapter 5, analysing data from disconnected trials should be avoided at all cost.

A hybrid can only be selected if it has been tested in at least three trials of the same MET. This condition assures that the genetic performance of each selected hybrid can be estimated with a reasonable accuracy, even if all its phenotypic records are obtained from single-replicate trials.

All selected inbred lines must be parent of at least two distinct hybrids. If a line is tested in only one hybrid, it is not possible to get a precise estimate of its GCA value.

All selected hybrids must have been evaluated in one or several METs that are fully connected. In theory, a set of METs is connected if each MET has at least one hybrid in common with another MET in the selection. In an attempt to improve the quality of the resulting phenotypic data, we enforce a slightly more stringent connectivity

criterion by selecting only METs that have at least two hybrids in common with one or two other METs.

These criteria have resulted in a selection of 105 ISSS lines and 92 Iodent lines resulting in 2361 single-cross, interheterotic hybrids. These hybrids only represent 24.4 % of all possible crosses in the theoretical half-diallel between the selected ISSS and Iodent lines. The phenotypic data on these hybrids is also severely unbalanced on a secondary level as each selected hybrid is on average present in 2.6 of a total of 1284 METs. These METs also contain phenotypic measurements for 33991 additional hybrids which have only one or no parental inbred lines in the selection. These hybrids are referred to as check varieties or checks and their measurements allow to connect the different METs. There are 209794 plots in this selection for which yield and grain moisture content measurements were recorded while days until flowering was only recorded for 148234 of them.

4.2 Molecular marker data

The 197 selected inbred lines were fingerprinted using microsatellite or simple sequence repeat markers (SSR) and amplified fragment length polymorphism markers (AFLP).

4.2.1 SSR

The selected set of inbred lines was fingerprinted by means of 101 SSR markers with known positions on the proprietary linkage map of RAGT R2n. This selection of markers is more or less evenly distributed over the 10 maize chromosomes. Due to problems identifying some SSR alleles (null alleles), only 75 markers have complete profiles over all selected inbred lines. In this dissertation, only the information of these complete SSR loci was used. 2.6% of all SSR locus/inbred line combinations is heterozygous, preventing an exact deduction of the hybrid genotype when these lines are used as parents. The average Polymorphism Information Content (PIC) of these 75 SSR loci is 0.55.

4.2.2 AFLP

The AFLP marker scores were generated according to the protocol of Vos et al. (1995) using 11 *Pst*I-*Mse*I and 4 *Eco*RI-*Mse*I primer combinations. The *Eco*RI and *Mse*I primers each had three selective nucleotides, while there were only two for the *Pst*I primers. There was a preference for the *Pst*I-*Mse*I primer combinations as the resulting markers are likely to be

Table 4.1: Overview of the 11 *Pst*I-*Mse*I and 4 *Ecor*I-*Mse*I primer combinations that were used to generate the AFLP-based fingerprints for the 197 selected inbred lines. The first row and column of the table give a specific name to each primer according to the naming scheme used by Vuylsteke et al. (1999). Primer names starting with M stand for *Mse*I-based primers, P for *Pst*I and E for *Ecor*I primers. The second row and column of the table contain the selective nucleotides which identify the specific primer. Primer combinations which were used to generate the AFLP fingerprints are indicated with an × while blank cells indicate primer combinations that were not used.

		M47	M48	M49	M50	M51	M55	M59	M61	M62
		CAA	CAC	CAG	CAT	CCA	CGA	CTA	CTG	CTT
P12	AC	×	×	×	×			×	×	×
P13	AGC	×	×	×				×		
E38	ACTC					×				
E39	AGAC						×	×		
E46	AGGC							×		

more evenly distributed over the maize genome than *Eco*RI-*Mse*I markers (Vuylsteke et al., 1999; Castiglioni et al., 1999). Table 4.1 gives an overview of the 15 primer combinations which produced 569 polymorphic bands for the 197 selected inbred lines.

CHAPTER 5

Graph-based data selection for the construction of genomic prediction models

5.1 Introduction

Despite the numerous studies devoted to molecular marker-based breeding, the genetic progress of most complex traits in today's plant and animal breeding programmes still heavily relies on phenotypic selection. Most breeding companies have established dedicated databases that store the vast number of phenotypic records that are being routinely collected throughout the course of their breeding programmes. These phenotypic records are, however, gradually being complemented by various types of molecular marker scores and it is to be expected that effective marker-based selection schemes will eventually allow to reduce current phenotyping efforts (Bernardo, 2008; Hayes et al., 2009). The available marker and phenotypic databases already allow for the construction and validation of marker-based selection schemes. Mining the phenotypic databases of a breeding company is, however, quite different from analysing the data that is generated by a carefully designed experiment. Genetic evaluation data is often severely unbalanced as elite genotypes are usually tested many times on their way to becoming a commercial variety or sire, while less performing genotypes are often disregarded after a single trial. Furthermore, the different phenotypic evaluation trials are separated in time and space and as such, subjected to different environmental conditions. Therefore, ranking the performance of genotypes that were evaluated in different phenotypic trials is usually a non-trivial task.

The content of this chapter has been submitted as Maenhout, S., De Baets B. and Haesaert G. (2009). Graph-based data selection for the construction of genomic prediction models. *Genetics*

Animal breeders are well experienced when it comes to handling unbalanced genetic evaluation data. The best linear unbiased predictor or BLUP approach (Henderson, 1975) presented a major breakthrough in this respect, especially when combined with restricted maximum likelihood or REML estimation of the needed variance components (Patterson and Thompson, 1971). Somewhat later on, this linear mixed modelling approach was also adopted by plant breeders as the *de facto* standard for handling unbalanced phenotypic data. The more recent developments in genomic selection (Bernardo, 1995; Meuwissen et al., 2001; Gianola and van Kaam, 2008) and marker-trait association studies (Yu et al., 2006) are, at least partially, BLUP-based and are therefore, in theory, perfectly suited for mining the large marker and phenotypic databases that back each breeding programme. In practice, however, the unbalancedness of the available genetic evaluation data often reduces its total information content and the construction of a marker-based selection model is limited to a more balanced subset of the data.

The most extreme case of an unbalanced design is a disconnected design. Table 5.1 gives an example of a disconnected sire evaluation design taken from Kennedy and Trus (1993). The breeding values of four sires are evaluated by measuring the performance of their offspring in three different herds. Sires having offspring in different herds provide vertical connections between herds while herds containing offspring of different sires provide horizontal connections. In a perfectly balanced design, each sire would have the same number of offspring tested in each herd. In the presented scenario however, sires s_1 and s_2 are disconnected from sires s_3 and s_4 as there is no possible path between these groups. This means that if we analyse the phenotypic data from this design with an ordinary least squares model, contrasts involving sires that belong to the disconnected groups would be inestimable. However, if we fit a linear mixed model to this data in which we assume herds as fixed and sires as random effects, contrasts involving sire BLUPs belonging to these disconnected groups are perfectly estimable. Ignoring connectivity issues by treating sire effects as random variables is, however, not without consequence. This approach implicitly assumes that all evaluated genotypes originate from the same population and as such have the same breeding value expectation. This assumption is generally not valid in animal breeding programmes as the better sires are usually evaluated in the better herds (Foulley et al., 1990). A similar stratification can be observed in genetic evaluation trials performed by plant breeders where late and therefore higher yielding genotypes are generally tested in geographical regions with longer growing seasons. As a consequence, BLUP-based genomic selection routines will be less efficient, while marker-trait association studies will suffer from increased false positive rates and reduced power. A very unbalanced but nevertheless connected design will also reduce the effectiveness of marker-based selec-

tion approaches as the prediction error variance of the estimated breeding values increases substantially. Furthermore, the estimated breeding values will be regressed towards the mean and will not account for the true genetic trend.

As phenotypic data is available, genotyping costs limit the total number of genotypes that can be included in the construction of a genomic prediction model. The best results will be obtained by selecting a subset of genotypes for which the phenotypic evaluation data exhibits the least amount of unbalancedness. In this paper we demonstrate how this phenotypic subset selection problem can be translated into a standard graph theory problem which can be solved with exact algorithms or less time consuming heuristics.

In most plant and animal species, the number of available molecular markers is rapidly increasing, while the genotyping cost per marker is decreasing. Nevertheless, as budgets are always limited, genotyping all mapped markers for a small number of genotypes might be less efficient than genotyping a restricted set of well-chosen markers on a wider set of genotypes. One should therefore be able to select a subset of molecular markers that covers the entire genome as uniformly as possible. We demonstrate how also this marker selection problem can be translated into a well-known graph theory problem which has an exact solution.

The third problem we tackle by means of graph theory is more specific to hybrid breeding programmes where the parental genotypes are nearly or completely homozygous. This implies that we can deduce the molecular marker fingerprint of a hybrid genotype from the marker scores of its parents. As the phenotypic data is collected on the hybrids, genotyping costs can be reduced by selecting a subset of parental inbreds that have produced the maximum number of genetically distinct offspring amongst themselves. Obviously, the phenotypic data on these offspring should be as balanced as possible.

Besides solving the above-mentioned selection problems by means of graph theory algorithms, we demonstrate their use in a simulation study that allows to determine the optimum trade-off between the number of genotypes and the size of the genotyped molecular marker fingerprint for predicting the phenotypic performance of a hybrid genotype by means of ε-insensitive support vector machine regression (ε-SVR) and Best Linear Prediction (BLP) (Bernardo, 1994, 1995, 1996a).

Table 5.1: Example of a disconnected sire × herd design taken from Kennedy and Trus (1993). The cell numbers indicate how many offspring of the sire pertaining to that particular column were evaluated in the herd pertaining to that particular row.

	s_1		s_2	s_3		s_4
h_1	3	—	6	0		0
	|		|			
h_2	3	—	4	0		0
h_3	0		0	7	—	5

5.2 Selecting genotypes from unbalanced phenotypic evaluation data

In most plant or animal breeding programmes, all phenotypic measurements that were recorded during genetic evaluation trials are stored for future reference. We can assume that this entire data set contains unbalanced phenotypic measurements on t genotypes, where t is generally a very large number. The available phenotypic data allows the breeder to try out one or more of the more recent BLUP-based genomic selection approaches without setting up dedicated trials. Given his financial limit for genotyping, he wants to select exactly p genotypes from this data set. The selection of p genotypes should be optimal in the sense that the precision of the BLUPs of the p breeding values that are obtained from a linear mixed model analysis of the full set of phenotypic records, is superior to the precision of any other set of BLUPs with cardinality p. This optimality criterion requires a measure of precision of a subset of BLUPs obtained from a linear mixed model analysis. To introduce this criterion, we will make the general assumption that the applied linear mixed model takes the form

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{Z}\boldsymbol{u} + \boldsymbol{e}\,, \tag{5.1}$$

where $\boldsymbol{y}$ is a column vector containing n phenotypic measurements on the t genotypes. $\boldsymbol{\beta}$ is a vector of fixed nuisance effects like trial, herd and replication effects and $\boldsymbol{u}$ is vector containing random genetic effects for each of the t genotypes. For ease of explanation, we

will assume that $\boldsymbol{u}$ only contains t breeding values, but the presented approach can easily be generalised to cases where $\boldsymbol{u}$ is made up from GCA and SCA effects and possibly the different levels of various G×E interaction factors. Vector $\boldsymbol{e}$ contains n random residuals. Matrices $\boldsymbol{X}$ and $\boldsymbol{Z}$ link the appropriate phenotypic records to the effects in $\boldsymbol{\beta}$ and $\boldsymbol{u}$ respectively. Furthermore we assume that we can represent the variance of $\boldsymbol{u}$ and $\boldsymbol{e}$ as

$$\mathrm{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{G} & \boldsymbol{0}\\ \boldsymbol{0} & \boldsymbol{R}\end{bmatrix}.$$

$\boldsymbol{G}$ can contain an assumed covariance structure for the t genotypes, typically a scaled numerator relationship matrix calculated from available pedigree or marker data. It is, however, important to realize that fitting a covariance between breeding values allows the BLUPs from genotypes which have little phenotypic information themselves, to borrow strength from phenotypic records on closely related genotypes. As a result, the p genotypes with the highest BLUP precision will most likely be close relatives which is detrimental for the generalising capabilities of the marker-based selection model. If we want the selection of p genotypes to rely completely on the amount of information and the structure (balancedness) of their phenotypic records, $\boldsymbol{G}$ should be a scaled identity matrix. Once the p genotypes have been selected, a pedigree or marker-based covariance structure can be incorporated in $\boldsymbol{G}$ for the construction of the actual marker-based prediction model. The covariance structure of the residuals in matrix $\boldsymbol{R}$ can contain heterogeneous variances for the different production environments, or in case that data originates from actual field trials, spatial information like row or column correlations. The BLUPs in vector $\boldsymbol{u}$ are obtained by solving the mixed model equations (Henderson, 1984)

$$\begin{bmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{bmatrix}\begin{bmatrix}\hat{\boldsymbol{\beta}}\\ \hat{\boldsymbol{u}}\end{bmatrix} = \begin{bmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{y}\\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{y}\end{bmatrix}.$$

The inverse of the coefficient matrix allows to obtain the prediction error variance (PEV) matrix of vector $\hat{\boldsymbol{u}}$ as

$$\begin{bmatrix}\boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{X}'\boldsymbol{R}^{-1}\boldsymbol{Z}\\ \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{X} & \boldsymbol{Z}'\boldsymbol{R}^{-1}\boldsymbol{Z}+\boldsymbol{G}^{-1}\end{bmatrix}^{-1} = \begin{bmatrix}\boldsymbol{C_{11}} & \boldsymbol{C_{12}}\\ \boldsymbol{C_{21}} & \boldsymbol{C_{22}}\end{bmatrix},$$

where

$$\mathrm{PEV}(\hat{\boldsymbol{u}}) = \mathrm{Var}(\hat{\boldsymbol{u}}-\boldsymbol{u}) = \boldsymbol{C_{22}}.$$

A logical choice of measure to express the precision of a selection of p BLUPs from the t candidates in vector $\hat{\boldsymbol{u}}$ would be some function of the $p \times p$ principal submatrix $\boldsymbol{C}_{\boldsymbol{22}}^{\boldsymbol{p}}$,

obtained by removing the rows and columns of $\boldsymbol{C_{22}}$ that pertain to genotypes that are not in that particular selection. As a good design is strongly associated with the precision of pairwise contrasts (Bueno and Gilmour, 2003), we use the lowest precision of all possible pairwise contrast vectors between the p selected genotypes as optimisation criterion. A pairwise contrast vector $\boldsymbol{q}^{ij}$ for the genotypes i and j is a vector where $q_i^{ij} = 1$ and $q_j^{ij} = -1$, while all other elements of $\boldsymbol{q}^{ij}$ are zeros. Laloé (1993) and Laloé et al. (1996) propose to express the precision of a linear contrast vector $\boldsymbol{q}$ by means of the generalised Coefficient of Determination which is defined as

$$\mathrm{CD}(\boldsymbol{q}) = \frac{\boldsymbol{q}'(\boldsymbol{G} - \boldsymbol{C_{22}})\boldsymbol{q}}{\boldsymbol{q}'\boldsymbol{G}\boldsymbol{q}}\,,$$

where $\mathrm{CD}(\boldsymbol{q})$ always lies within the unit interval. They indicate that $\mathrm{CD}(\boldsymbol{q})$ can be obtained as a weighted average of the $t-1$ non-zero eigenvalues μ_i of the generalised eigenvalue problem

$$((\boldsymbol{G} - \boldsymbol{C_{22}}) - \mu_i \boldsymbol{G})\boldsymbol{v}_i = \boldsymbol{0}\,, \tag{5.2}$$

as

$$\mathrm{CD}(\boldsymbol{q}) = \frac{\sum_{i=2}^{t} a_i^2 \mu_i}{\sum_{i=2}^{t} a_i^2}\,, \tag{5.3}$$

where the first eigenvalue μ_1 always equals zero as a consequence of the well-known summation constraint $\boldsymbol{1}'\boldsymbol{G}^{-1}\hat{\boldsymbol{u}}$ (see e.g. Foulley et al. (1990)). Each linear contrast vector $\boldsymbol{q}$ can be expressed as a linear combination of the $t-1$ non-zero eigenvectors $\boldsymbol{v}_i$ as

$$\boldsymbol{q} = \sum_{i=2}^{t} a_i \boldsymbol{v}_i\,.$$

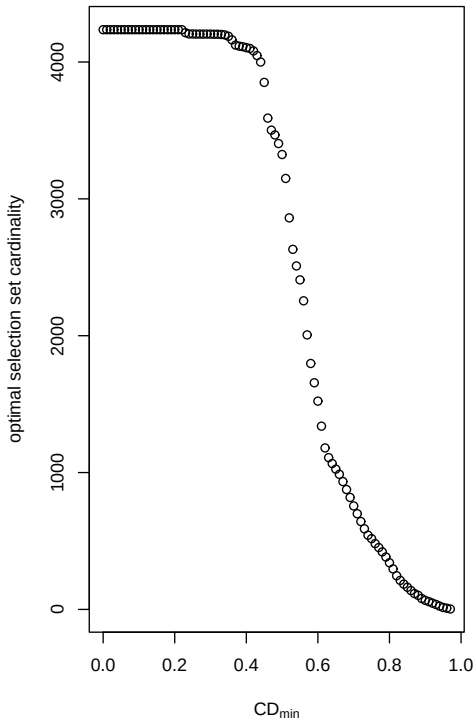
In fact, all linear contrast vectors that are estimable in the least squares sense, are linear combinations of the eigenvectors $\boldsymbol{v}_i$ of Eq. (5.2) that are associated to non-zero eigenvalues μ_i, while those contrasts that are not estimable in a least squares sense, are linear combinations of eigenvectors for which at least one associated eigenvalue is zero. This implies that the CD of a pairwise contrast vector involving two genotypes that were evaluated in two disconnected groups does not necessarily become zero as several eigenvalues μ_i in Eq. (5.3) might be non-zero. This might bias the selection procedure to favour a disconnected set of genotypes with a high information content (i.e. a high level of replication) instead of a connected set of genotypes with low information content. To avoid this situation, the CD of pairwise contrast vectors between disconnected genotypes should be forced to zero. In case

Eq. (5.1) represents a simplified animal model where $\boldsymbol{G} = \boldsymbol{I}\sigma_g^2$ and $\boldsymbol{R} = \boldsymbol{I}\sigma_e^2$, disconnected pairs of genotypes can be easily identified by examining the block diagonal structure of the PEV matrix $\boldsymbol{C_{22}}$ as explained in Appendix A1 of this chapter. In Appendix A2 we show how disconnected genotype pairs can be identified by means of the transitive closure of the adjacency matrix of the t genotypes.

Now that we have the corrected CD for each of the $\frac{p(p-1)}{2}$ pairwise contrast vectors, we can represent the t genotypes as vertices (also called nodes) of a weighted complete graph where the edge between genotype i and genotype j carries the weight $\mathrm{CD}(\boldsymbol{q}^{ij})$, expressing the precision of the pairwise contrast as a number between zero and one. We need to select exactly p vertices such that the minimum edge weight in the selected subgraph is maximised. This problem is equivalent to the 'discrete p-dispersion problem' from the field of graph theory. This problem setting is encountered when locating facilities that should not be clustered in one location, like nuclear plants or franchises belonging to the same fast-food chain. This problem is NP-hard even when the distance matrix satisfies the triangle inequality. Erkut (1990) describes two exact algorithms based on a branch and bound strategy and compares 10 different heuristics (Erkut et al., 1994). An interesting solution lies in the connection between the discrete p-dispersion problem and the maximum clique problem. A clique in a graph is a set of pairwise adjacent vertices, or in other words, a complete subgraph. The corresponding optimisation problem, the maximum clique problem, is to find the largest clique in a graph. This problem is also NP-hard (Carraghan and Pardalos, 1990). The idea is to decompose the discrete p-dispersion problem in a number of maximum clique problems by assigning different values to the minimum required contrast precision $\mathrm{CD}_{\mathrm{min}}$. Initially, $\mathrm{CD}_{\mathrm{min}}$ is low (e.g. $\mathrm{CD}_{\mathrm{min}} = 0.1$) and we define a graph $G'(V, E')$ where the edges of the original graph G are removed when their edge weight is smaller than $\mathrm{CD}_{\mathrm{min}}$. This implies that there will be no edges between disconnected genotype pairs in the derived graph as these edge weights have been set to zero by the CD correction procedure. Solving the maximum clique problem in $G'(V, E')$ allows to identify a complete subgraph for which all edge weights are guaranteed to be greater than $\mathrm{CD}_{\mathrm{min}}$. The number of vertices in this complete subgraph is generally smaller than t but greater than p. By repeating this procedure with increasing values of $\mathrm{CD}_{\mathrm{min}}$ one can make a trade-off between sample size and sample quality as is demonstrated in Figure 5.1 for a representative sample of size $t = 4236$ genotypes for which of the genetic evaluation data was recorded as part of the grain maize breeding programme of the private company RAGT R2n. Each dot represents the largest possible selection of genotypes where $\mathrm{CD}_{\mathrm{min}}$ ranges from 0 to 0.97. The data used in this example is connected as there is no sudden drop in the number of genotypes when $\mathrm{CD}_{\mathrm{min}}$ is raised from 0.0 to 0.1. In general, the surface below the curve represents

a measure of data quality. If one is only interested in obtaining the optimal selection of exactly p genotypes from a set of t candidates, one can implement the described maximum clique-based procedure in a binary search.

Figure 5.1: Graphical representation of the trade-off between the selection size and the selection quality for a sample of the RAGT grain maize breeding pool. For each examined level of CD_{min}, ranging from 0.0 tot 0.97, the dot represents the maximum cardinality selection of genotypes for which the minimum precision of a pairwise contrast is at least CD_{min}.



The presented approach for solving the discrete p-dispersion problem requires an efficient algorithm to obtain the maximum clique from a graph. Several exact algorithms and heuristics have been published, but comparing these is often difficult as the dimensions and densities of the provided example graphs as well as computational platforms tend to differ between papers. The exact algorithm of Carraghan and Pardalos (1990) is, however, considered as the basis for most later algorithms. Although the efficiency of this algorithm has been superseded by that of more recent developments (Östergård, 2002; Tomita and

Seki, 2003), its easy implementation often makes it the method of choice. If the available run-time is limited, a time-constrained heuristic like the Reactive Local Search approach presented by Battiti and Protasi (2001) might be more appropriate. Bomze et al. (1999) give an overview of several other heuristic approaches found in literature, in particular greedy construction and stochastic local search including simulated annealing, genetic algorithms and tabu search.

5.3 Selecting markers from a dense molecular fingerprint

The construction of a genomic prediction model requires genotypic information on each of the p selected genotypes. Generally it is assumed that a good prediction accuracy can only be achieved by maximising the genome coverage, which implies genotyping a large number of molecular markers. This approach seems particularly attractive as genotyping costs are decreasing rapidly. However, as will be shown in Chapter 8, the relation between the number of genotyped markers and the obtained prediction accuracy seems to be subject to the law of diminishing marginal returns. This means that it might be more efficient to construct the genomic prediction model using a larger number of genotypes in combination with a smaller molecular fingerprint. This subset of molecular markers should cover the genome as uniformly as possible such that the probability of detecting a marker-trait association is maximised.

We start by solving this selection problem on a single chromosome for which t candidate molecular markers have been mapped. We want to select exactly q of these markers such that the chromosome coverage is optimal compared to all other possible selections of q markers. Maximising the chromosome coverage could mean several things, including maximising the average intermarker distance and maximising the minimum marker distance. We prefer the latter definition as it implies a one-dimensional version of the discrete p-dispersion problem. In this restricted setting, a reduction to a series of maximum clique problems is not necessary as Ravi et al. (1991) have published an algorithm that obtains the optimal solution in an overall running time of $\mathrm{O}(\min(t^2, qt\,\log(t)))$.

The extension to $c > 1$ chromosomes is again dependent on the interpretation of a uniform genome coverage. In case this means an approximately equal number of uniformly distributed markers on each chromosome we can select $\frac{q}{c}$ markers on each chromosome using the above-mentioned algorithm. If q is not an exact multiple of c, the remainder after division could be attributed to each of the different chromosomes in decreasing order of their minimum intermarker distance after the addition of one marker. A more intuitive

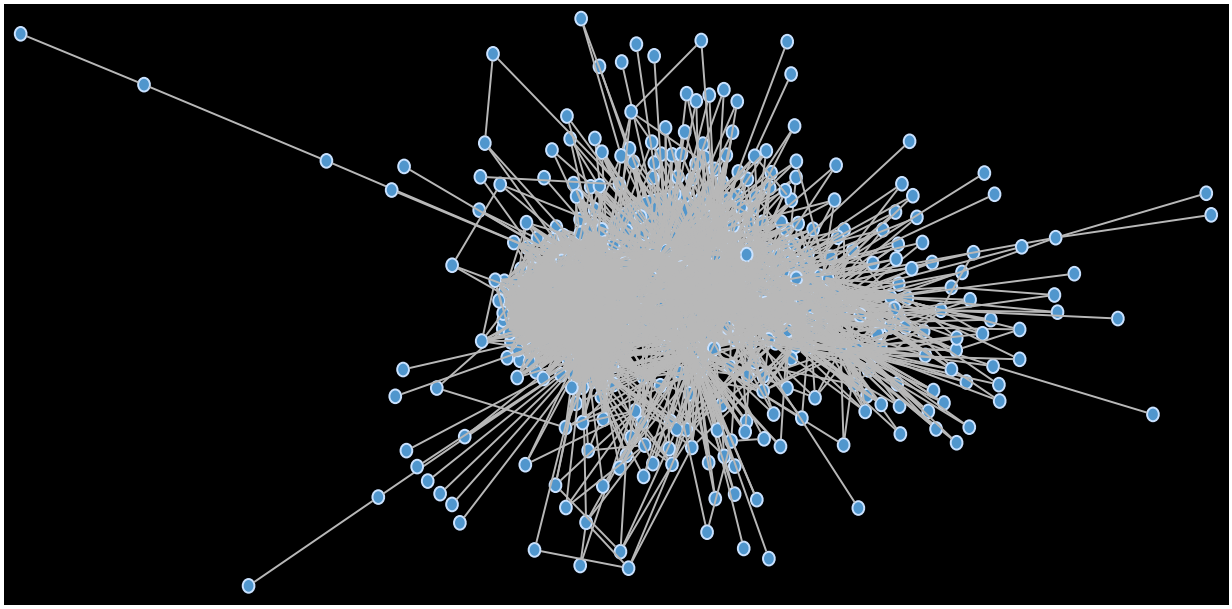
interpretation of a uniform genome coverage entails a selection of markers such that the minimum intermarker distance over all chromosomes is maximised. This can be achieved by linking all chromosomes head to tail as if all markers were located on a single chromosome. To be able to use the above-mentioned algorithm, the distance between the last marker of the first chromosome and the first marker of the second chromosome of each linked chromosome pair should be set to infinity.

5.4 Selecting parental inbred lines

In hybrid breeding programmes, the molecular marker fingerprint of a single-cross hybrid can be easily deduced from the fingerprints of its two homozygous parents. This allows to reduce the total genotyping cost of the genomic prediction model considerably. If we assume we have a budget for fingerprinting exactly k parental inbred lines, we can maximise the number of genotyped single-cross hybrids by selecting the set of lines which have produced the maximum number of single-cross hybrids amongst themselves. We approach this selection problem by representing the total set of parental inbred lines as the vertices of an unweighted pedigree graph where an edge between two vertices represents an offspring genotype (i.e. a single-cross hybrid) for which genetic evaluation data is available. Figure 5.2 shows such a graph representation of the sample used in Figure 5.1 containing 487 inbred lines and 4236 hybrids. We need to select a k-vertex subgraph which has the maximum number of edges. In graph theory parlance, this problem is called the 'densest k-subgraph problem' which is shown to be NP-hard. Several approximation algorithms have been published including the heuristic based on semidefinite programming relaxation presented by Feige and Seltser (1997) and the greedy approach of Asahiro et al. (2000). The basic idea of the latter is to repeatedly remove the vertex with minimum degree (i.e. minimum number of edges) from the graph until there are exactly k vertices left. This approach has been shown to perform almost just as good as the much more complicated alternative based on semidefinite programming.

The presented selection procedure does not consider the quality of the genetic evaluation data that is available for the hybrids. As a result, the optimal selection with respect to the maximisation of the number of training examples might turn out to be a very poor selection with respect to the quality of the phenotypic data. To enforce these data quality constraints, the described inbred line selection procedure should be performed after a preselection of the hybrids based on the precision of pairwise contrasts. If we select k inbred lines where k ranges from the total number of candidate parents to 3 for each level of

Figure 5.2: Graph representation of a sample of the RAGT grain maize breeding pool. The vertices represent inbred lines and the edges are single-cross hybrids.

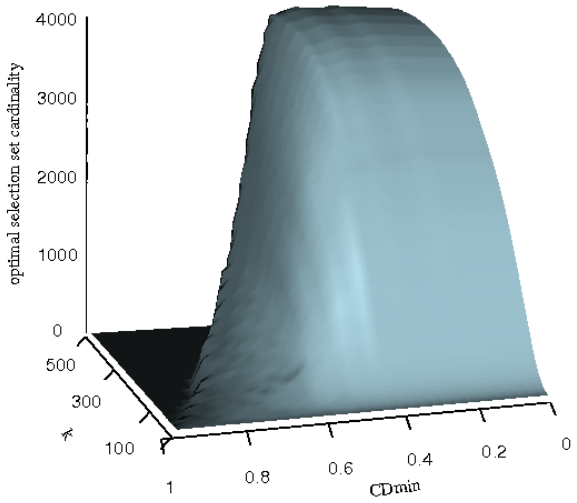


CD_{min} ranging from 0.1 tot 0.97 we get a 3-dimensional representation of the data quality as shown in Figure 5.3. Similarly to Figure 5.1, each dot on the surface represents the size of the optimal selection of hybrids under the constraints of a genotyping budget for k parental inbred lines and a minimum pairwise contrast precision of CD_{min}.

5.5 Simulation study

The construction of a hybrid prediction model based on ε-insensitive support vector machine regression (ε-SVR) (Chapters 7 and 8) or Best Linear Prediction (BLP) (Bernardo, 1995) requires a combination of genotypic and phenotypic data on a predefined number of inbred lines and their hybrid offspring respectively. As phenotypic data is available from past genetic evaluation trials, the number of training examples that is used for the construction of this prediction model is constrained by the total genotyping cost. If we reduce the size of the fingerprint, more inbred lines can be genotyped and more training examples become available which should result in a better prediction accuracy of the model. However, reducing the size of the molecular marker fingerprint comes at the price of a reduced genome coverage and an increased number of selected hybrids results in a reduced

Figure 5.3: Graphical representation of the trade-off between the selection size and the selection quality when only k parental inbred lines are being genotyped. For each examined level of CD_{min} ranging from 0.0 tot 0.97 the number of genotyped inbred lines k is reduced from 487 tot 3. Each dot in the plotted surface represents the maximum cardinality selection of hybrid genotypes for which the minimum precision of a pairwise contrast is at least CD_{min} and the number of parents is exactly k.



precision of BLUP contrasts due to connectivity issues (e.g. Figure 5.1). Therefore, it is to be expected that within the constraints of a fixed genotyping budget, maximum prediction accuracy can be achieved by finding the optimal balance between the fingerprint size and the number of training examples. The location of this optimum is obviously highly dependent on the information content of the available phenotypic data and the applied linkage map, but can be estimated by means of the afore mentioned graph theory algorithms for each specific data set.

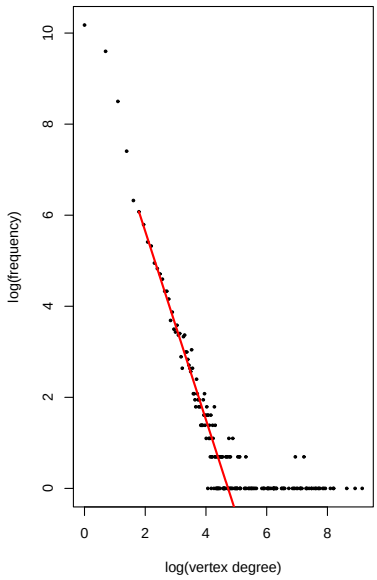
5.5.1 Simulation setup

To demonstrate the approach, we use the phenotypic data that was generated as part of the grain maize breeding programme of the private breeding company RAGT R2n and their proprietary SSR linkage map. We assume to have a limited budget for genotyping 101 SSR markers on 200 inbred lines or 20200 markers in total. We will also assume that we can limit the number of candidate inbred lines to 400 by restricting the prediction model to a specific heterotic group combination, a specific environment (i.e. maturity rating, irrigation and fertiliser treatments, ...) and the set of inbred lines that are available at the moment of genotyping. This intensive preselection of candidate lines is mainly needed for keeping the simulations tractable. In a more realistic setting, calculations are only performed once so the set of initial candidate lines can be larger. Table 5.2 gives a schematic overview of the different steps that are performed at each iteration of the simulation routine.
Again we make use of the pedigree graph representation where inbred lines are represented as vertices and each single-cross hybrid is represented as an edge between two vertices as shown in Figure 5.2. In this graph, the degree of a vertex (i.e. the number of edges incident to the vertex) therefore equals the number of distinct single-cross hybrids of which the inbred line is a parent. Figure 5.4 shows the empirical distribution of these degrees on a log scale for the entire RAGT grain maize breeding pool. The observed long-tailed behaviour of the empirical distribution is not unexpected as most inbred lines only have a limited number of children, while inbred lines with higher progeny numbers (i.e. the tester lines) are rare. In an attempt to parametrise the underlying distribution from which the observed vertex degrees were drawn, several candidate distributions among which the Poisson, geometric, discrete log-normal and discrete power-law distributions were fitted by means of likelihood maximisation. The best fit was observed for the discrete power-law distribution with a left threshold value of 6 which is indicated as a straight line on Figure 5.4. The fit of this distribution is, however, insufficient as indicated by the significantly large Kolmogorov-Smirnov D-statistic where significance is determined by means of the parametric bootstrap

Table 5.2: Description of each step that is performed during a single iteration of the simulation routine. The goal is to find the optimal trade-off between the number of genotyped inbred lines and the size of their molecular fingerprint, when the total genotyping budget is fixed.

step	description
1	sample 400 vertices from the pedigree graph by means of the 'Forest Fire' algorithm: $\Rightarrow$ indirect sampling of hybrids $\Rightarrow$ indirect sampling of METs
2	partition sampled inbred lines in c heterotic groups by means of the Dsatur vertex colouring algorithm
3	simulate 8 breeding cycles on each of the c heterotic groups
4	simulate phenotypic records on the sampled hybrids
5	reduce the number of sampled hybrids by gradually increasing $\mathrm{CD}_{\mathrm{min}}$
6	reduce the number of genotyped inbred lines by means of the greedy densest k-subgraph algorithm
7	select q SSR markers with maximal genome coverage
8	determine the prediction accuracy of ε-SVR and BLP using the reduced set of training examples

Figure 5.4: Log-scaled degree distribution of the graph created from part of the RAGT R2n grain maize breeding programme. In this undirected, unweighted graph, parental inbred lines are represented as vertices and single-cross hybrids as edges. Each dot represents a unique log-scaled vertex degree (horizontal axis) and the log of its frequency in the graph (vertical axis). The straight line represents the fitted power law distribution by means of likelihood maximisation. The threshold value of 6 was determined by minimising the Kolmogorov-Smirnov statistic as described by Clauset et al. (2009).



procedure described by Clauset et al. (2009).

As no conclusive evidence on the underlying distribution of the observed vertex degrees was found, we prefer to sample inbred lines from the full RAGT graph directly. However, taking a representive sample from a large graph is not a trivial task. The sample quality of various published graph sampling algorithms seems to be highly dependent on the properties of the graph. To decide which sampling routine is optimal for the RAGT data, we first need to decide on a measure of sample quality. We compare the empirical cumulative distribution (ECD) of the vertex degrees in the full graph with those ECDs of 100 samples containing 400 vertices. From these ECDs, we calculate the average Kolmogorov-Smirnov D-statistic for each examined sampling routine. For the RAGT data, the 'Forest Fire' vertex sampling approach resulted in the smallest average D-statistic compared to the alternative methods

described by Lescovec and Faloutsos (2006). This sampling routine starts by selecting a vertex v_0 uniformly at random from the graph. Vertex v_0 now spreads 'the fire' to a random selection of its neighbours which are then in turn allowed to infect a random selection of their own neighbours. This process is continued until exactly 400 vertices are selected. If the fire dies out before the sample is complete, a new starting vertex is selected uniformly at random. The number of neighbours that is infected at each selected vertex, is obtained as a random draw from a geometric distribution where the parameter p was set to 0.62, as this value resulted in the best average sample quality. All hybrids for which both parents were sampled (i.e. the edges of the subgraph) have associated phenotypic records and as such indirectly sample a set of multi-environment trials (METs). All hybrids that were not indirectly selected by the inbred line sample, but do have phenotypic records in the sample of METs, are included in the selection as data connecting check varieties. Despite the fact that the RAGT data already provides phenotypic records for the selected hybrids and check varieties, we replace these by simulated measurements as we want to be able to assess the actual prediction accuracy of ε-SVR and BLP under various levels of data quality.

The simulation of these phenotypic records for the sampled hybrids starts by partitioning the selected inbred lines into heterotic groups. This partitioning should ensure that the two parents of each single-cross hybrid always belong to distinct heterotic groups, while the total number of groups needs to be minimised. The graph theory equivalent of this problem is called the 'vertex colouring problem' which, as all previously described graph theory problems, belongs to the complexity class of NP-hard problems. The minimum number of colours (i.e. heterotic groups) is called the chromatic number of the graph. The vertex colouring problem has been extensively studied in graph theory literature (Jensen and Toft, 1995) and several efficient heuristics are available. The greedy desaturation algorithm or Dsatur published by Brélaz (1979) is often used as a benchmark method to assess the efficiency and precision of newly developed vertex colouring algorithms. Its good performance on a variety of graphs and easy implementation makes it the method of choice for designating inbred lines to heterotic groups at each iteration of the simulation routine. Once the chromatic number c has been determined for the sampled set of inbred lines, an entire breeding programme is simulated starting from c open-pollinated varieties and resulting in c unrelated heterotic groups. The simulation of this breeding programme mimics the maize breeding programme of the university of Hohenheim as described by Stich et al. (2007) and described in detail in Chapter 6. In short, the simulation routine uses the proprietary linkage map of the breeding company RAGT R2n containing 101 microsatellites and adds an additional 303 evenly distributed, simulated SSRs. It also

generates 250 QTL loci of the selection trait (e.g. yield) which are randomly positioned on the genetic map. The number of alleles for each SSR or QTL is drawn from a Poisson distribution with an expected value of 7. Each simulation starts by generating an initial base population in Hardy-Weinberg equilibrium. Allele frequencies for each locus are drawn from a Dirichlet distribution and used to calculate the allele frequencies in each of the c subpopulations assuming an Fst value of 0.14. We perform 8 breeding cycles where each cycle consists of 6 generations of inbreeding and subsequent phenotypic selection based on line *per se* or testcross performance as described by Stich et al. (2007). The result is a set of 400 highly selected inbred lines partitioned in c unrelated heterotic groups. Within each of these groups, the simulated inbred lines are randomly assigned to the sampled inbred lines and a genotypic value is generated for each interheterotic hybrid by summing the effects of the 250 QTL alleles of both parents and adding a normally distributed SCA value. The size of the SCA variance component depends on the heritability of the trait under consideration, but is assumed to be only $\frac{1}{8}$ of the total non-additive variance (SCA + G×E and residual error) as this was the average of observed ratios for the traits grain yield, grain moisture contents and days until flowering in the actual RAGT data. The genotypic values of the check varieties are generated from a single normal distribution where the variance is the sum of the additive variance and SCA variance of the sampled hybrids. The simulated genotypic values of hybrids and check varieties are used to generate phenotypic records according to the sampled MET data structure, assuming a single replication in each location of a MET. This implies that G×E effects are confounded with the residual error and only a single effect is drawn from a normal distribution where the variance is $\frac{7}{8}$ of the total non-additive variance. The main environmental effect of each location is also drawn from a normal distribution for which the variance is twice the additive variance of the hybrids.

The simulated phenotypic records that are associated with the sampled data structure allow to estimate the genotypic value of each hybrid by means of a linear mixed model analysis. We fit genotypes (hybrids and check varieties) as random and locations as fixed effects as this approach should result in an ε-SVR model with a superior prediction accuracy as will be shown in Chapter 8. To avoid selections of closely related hybrids, the variance-covariance matrix of the genotypic effects is fitted as a scaled identity matrix. The resulting PEV matrix of the random genotypic effects is used to iteratively select a smaller subset of the sampled hybrids by gradually increasing the minimum required precision of each pairwise contrast in the selection. Initially, the required CD_{min} value is set to 0 which implies that all hybrids are selected. The next examined level of precision requires $\text{CD}_{\text{min}} > 0$ which effectively excludes selections containing disconnected genotypes. More stringent

levels of precision are enforced by requiring $CD_{min} > q_p$ where q_p is the pth quantile of the observed distribution of CD values in the complete sample and p ranges from 0 tot 0.875 in steps of 0.125. Defining CD_{min} values as quantiles allows to compare the obtained prediction accuracies over the different samples of the simulation routine.
For each level of CD_{min}, the number of genotyped inbred lines is reduced from 400 to 50 in steps of 50, while at the same time the number of markers in the molecular fingerprint is increased from 50 to 404. For each combination of CD_{min} and number of genotyped inbred lines, the BLUPs of the selected hybrids are used to construct an ε-SVR and a BLP-based prediction model. In fact, the prediction accuracy of both methods is verified by randomly assigning the BLUPs to one of five groups. For each of these groups, a separate ε-SVR and BLP prediction model is constructed using all BLUPs in the remaining four groups as training data. The resulting prediction model is then used to make predictions on the hybrids in the selected group (i.e. the validation data). Combining the predictions of all five models allows to obtain a measure of prediction accuracy by correlating them against the simulated genotypic values.

5.5.2 Simulation results

We expect that enforcing a minimum required pairwise contrast precision $CD_{min} > 0$, results in a selection of BLUPs that has greater accuracy compared to the full set of hybrids. In Figure 5.5 this BLUP accuracy is plotted against CD_{min} and the maximum number of inbred lines for each of the three examined heritability levels. Each point on these wireframe surfaces represents the squared Pearson correlation between the BLUPs and the actual, simulated genotypic values of the selected hybrids at that particular level of CD_{min} and number of parental inbred lines, averaged over 100 iterations of the simulation routine. We can see that an increase in CD_{min} results in an almost linear increase in BLUP precision for each heritability level. This effect is especially pronounced for the lowest heritability level $h^2 = 0.25$. As expected, the BLUP precision is not influenced by the number of parental inbred lines.
Figure 5.6 presents the prediction accuracy of both ε-SVR and BLP for increasing values of the minimum required contrast precision CD_{min} and a decreasing number of genotyped inbred lines. The height of each point in the wireframes represents the average prediction accuracy, expressed as a squared Pearson correlation, over 100 iterations of the simulation routine. For each of the examined heritability levels, ε-SVR generally performs better than BLP. The negative effect of disconnected hybrids in the selection of training examples is visualised as the sharp increase in prediction accuracy when the minimum required contrast

Figure 5.5: Accuracy of the genotypic value BLUPs of the hybrids selected using the described graph-based procedures. The three examined heritability levels $h^2 = 0.25$, $h^2 = 0.5$ and $h^2 = 0.75$ are represented by the bottom, middle and top wireframe surfaces respectively. Each point on a surface is the squared Pearson correlation between the BLUPs and the actual (simulated) genotypic values of the selected hybrids under the constraints of a minimum required contrast precision $\mathrm{CD}_{\mathrm{min}}$, expressed as a percentile of the sampled CD values, and the number of genotyped inbred lines, averaged over 100 iterations of the simulation routine.

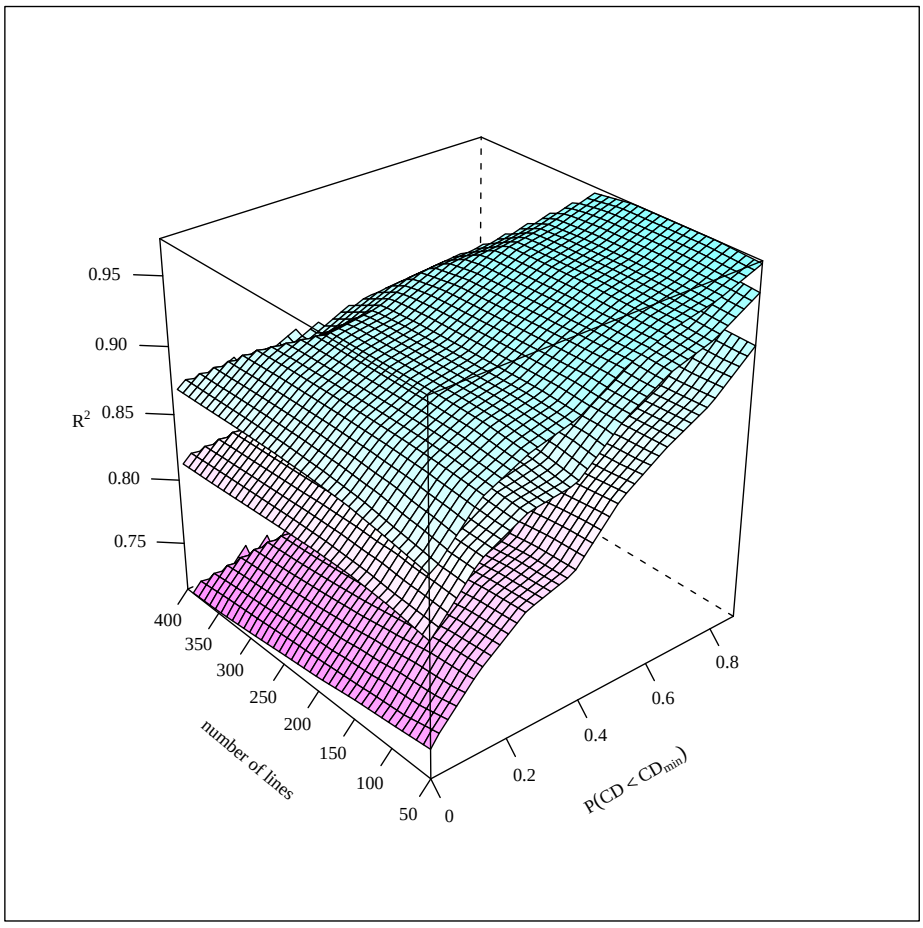
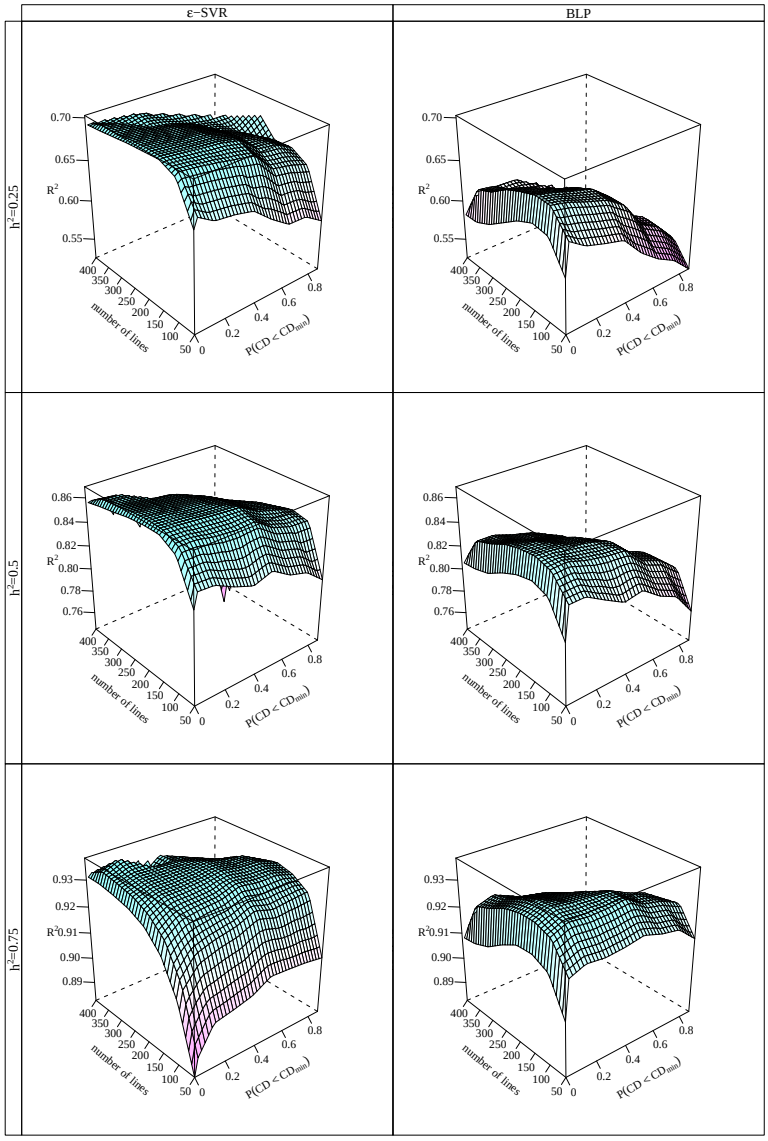


Figure 5.6: Average prediction accuracy of ε-SVR and BLP prediction models over 100 iterations of the simulation routine for varying levels of the minimum required contrast precision CD_{min}, expressed as a percentile of the sampled CD values ranging from 0 tot 0.875, and the number of genotyped inbred lines. The height of each point in the wireframe represents the prediction accuracy obtained by ε-SVR and BLP when training on the optimal selection of hybrids under the constraints imposed by the levels of the two independent variables. Prediction accuracy is expressed as the average squared Pearson correlation between the simulated and the predicted genotypic values of the hybrids. The scales of the vertical axes are only comparable within the same heritability level.



precision is slightly constrained from $CD_{min} = 0$ to $CD_{min} > 0$. This effect is more pronounced for BLP than for ε-SVR. Increasing CD_{min} any further, generally decreases the prediction accuracy, especially for traits with lower heritability. This observation implies that, at least for the RAGT data set, a larger number of training examples of lower data quality is to be preferred over a smaller selection of hybrids for which more and better connected phenotypic information is available, as long a disconnected genotypes are excluded.

BLP and ε-SVR do not take an unanimous stand on the optimal number of genotyped inbred lines. For BLP, the optimum seems to lie somewhere around 100 inbred lines for $h^2 = 0.25$ and 150 for $h^2 = 0.50$ and $h^2 = 0.75$, the equivalent of fingerprint sizes of 202 and 134 SSR markers respectively. This optimum is, however, less pronounced for the higher heritability levels. For ε-SVR, the optimal number of inbred lines is 150, 200 and 350 for $h^2 = 0.25$, $h^2 = 0.5$ and $h^2 = 0.75$ respectively. At the highest heritability level, ε-SVR seems to prefer training sets of maximum size, at the cost of a very small molecular fingerprint size. The observed behaviour of both BLP and ε-SVR is consistent with the results presented in Chapter 8, where it is shown that BLP is less sensitive to a reduction of the number of training examples compared to ε-SVR, as long as the molecular fingerprint is dense. ε-SVR on the other hand, although requiring a training set of considerable size, handles smaller or less informative molecular fingerprints better than BLP.

5.6 Discussion

This article presents three selection problems that are relevant to the budget-constrained construction of a genomic prediction model from available genetic evaluation data. The first problem considers the selection of exactly p genotypes from a set of t candidates that will be genotyped to serve as training examples for the construction of the prediction model. This selection should be optimal in the sense that a linear mixed model analysis of the associated phenotypic records should result in a set of p BLUPs of genotypic values that have the highest precision of all possible selections. By defining the precision of a selection as the minimum generalised coefficient of determination of a pairwise contrast, this selection problem can be translated to the 'discrete p-dispersion problem' from the field of graph theory. The reduction of this problem to a set of maximum clique problems allows to visualise the trade-off between selection size and selection quality. The greedy nature of a breeding programme does unfortunately bias the presented selection approach towards high-performing genotypes. These are generally tested more thoroughly than their

low-performing colleagues. As the latter generally only have a few associated phenotypic records, the pairwise contrasts involving these genotypes have a low precision which in turn makes their selection by the described procedure very unlikely. As a consequence, the resulting genomic prediction model is likely to overestimate the capabilities of the low-performing genotypes. To avoid this bias, the selection procedure should optimise two objectives simultaneously: (1) maximising the minimum precision of all pairwise contrasts in the selection and (2) maximising the genetic variance in the selection. Even if one would succeed in finding an acceptable trade-off between these conflicting objectives, the estimates of the genotypic value of low-performing genotypes will always suffer from large standard errors which makes them unreliable training examples.

The second problem we discuss deals with the selection of exactly q molecular markers from a set of t candidates for which the relative positions on a genetic map are known. To guarantee that the selection has an optimal genome coverage, we maximise the minimum intermarker distance. We show that this problem can be translated to a one-dimensional discrete p-dispersion problem for which an exact algorithm is available.

The third problem is specific to hybrid breeding programmes and entails the selection of exactly k parental inbred lines such that the number of single-cross hybrids in the selection is maximised. If we represent the inbred lines as vertices of a graph and each single-cross hybrid as an edge between its parental vertices, this problem can be translated to the 'densest k-subgraph problem' which we solve by using a greedy heuristic.

The presented solutions to the three selection problems are put into practice in a simulation study where the goal is to find the optimal number of training examples for the construction of ε-SVR and BLP prediction models with maximal prediction accuracy under a fixed genotyping budget. At each iteration of the simulation routine, inbred lines, hybrids and their associated phenotypic data structure are sampled from actual genetic evaluation data. The number of training examples is gradually reduced by putting constraints on the data quality and the number of genotyped inbred lines. The results indicate that selections of training examples containing disconnected genotypes are detrimental for the prediction accuracy of both ε-SVR and BLP. More stringent data quality constraints are however not necessary. ε-SVR performs best if the number of parental inbred lines (i.e. the number of training examples) is maximised at the cost of a reduced genome coverage. BLP on the other hand performs best when trained on a smaller set of training examples for which a dense fingerprint is available.

5.7 Appendix: Identifying disconnected pairs of genotypes

5.7.1 Examination of the PEV matrix

If we analyse the available genetic evaluation data with a linear mixed model according to Eq. (5.1) where the variance structure is simplified to

$$\mathrm{Var}\begin{bmatrix}\boldsymbol{u}\\ \boldsymbol{e}\end{bmatrix} = \begin{bmatrix}\boldsymbol{I}\sigma_g^2 & \mathbf{0}\\ \mathbf{0} & \boldsymbol{I}\sigma_e^2\end{bmatrix},$$

we can express the prediction error variance matrix as (Henderson, 1984)

$$\mathrm{PEV}(\hat{\boldsymbol{u}}) = (\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z} + \boldsymbol{I}\lambda)^{-1}\sigma_e^2\,,$$

where $\lambda = \frac{\sigma_e^2}{\sigma_g^2}$ and $\boldsymbol{M}$ is the orthogonal projector on the column space of matrix $\boldsymbol{X}$ as $\boldsymbol{M} = \boldsymbol{I} - \boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'$. The matrix product $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ is in fact the information matrix of the genotypic effects if we would consider both environments and genotypic effects as fixed and analyse the data as a block design in a linear least squares setting. Chakrabarti (1964) proves that if such a block design is fully connected (i.e. all elementary contrasts are estimable in a least squares sense), the rank of this information matrix equals $t-1$, where t is the number of fitted genotypic effects. Furthermore, Heiligers (1991) proves that in case the information matrix has a lower rank $t-p$, where $p \geq 2$, the design is disconnected and the symmetric matrix $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ can always be put in a block diagonal form with p distinct blocks around the principal diagonal, by simply permuting the appropriate rows and columns. Each of these blocks represents a set of fully connected genotypes that are disconnected from all other genotypes that are not represented in that particular block. If we assume that we are dealing with a disconnected design and that the columns of $\boldsymbol{Z}$ (i.e. the genotypes) are ordered in such a way that $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}'$ is in block diagonal form, it should be fairly obvious that also $\mathrm{PEV}(\hat{\boldsymbol{u}})$ is block diagonal as inversion preserves this matrix property. As most linear mixed model packages provide the PEV matrix, the identification of disconnected genotype pairs can be performed by recovering this block diagonal structure by appropriate row and column permutations.

5.7.2 Transitive closure

If $\mathrm{Var}(\boldsymbol{u})$ is not a diagonal matrix, the block diagonal structure of $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ is not preserved in the PEV matrix. One could of course examine the structure of $\boldsymbol{Z}'\boldsymbol{M}\boldsymbol{Z}$ instead, but

this matrix is generally not available. It might therefore be easier to identify disconnected genotype pairs by determining the transitive closure of their adjacency matrix. This is a symmetric, Boolean $t \times t$ matrix where the element on row i and column j is set to 1 if genotypes i and j have been evaluated in a common environment and 0 otherwise. For the example in Figure 5.1, this adjacency matrix looks like

$$\boldsymbol{A} = \begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{bmatrix}$$

The transitive closure of this matrix is again a symmetric, block-diagonalisable, Boolean matrix which can be interpreted in a similar way as $\boldsymbol{Z'MZ}$. Warshall (1962) describes a concise and efficient algorithm for computing the transitive closure of an adjacency matrix which has a worst case complexity of $\mathrm{O}(t^3)$. More advanced algorithms are described by Naessens et al. (2002) and De Meyer et al. (2004).

CHAPTER 6

Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes

6.1 Introduction

The coefficient of coancestry (CoC) between two individuals i and j is defined as the probability that at an allele drawn from both i and j at the same locus is identical by descent (IBD) from a recent common ancestor. This similarity measure is frequently used for modelling the covariance between the genetic background of plants involved in breeding programmes (Panter and Allen, 1995a,b; Bernardo, 1994, 1995, 1996a,c) or association studies (Jannink et al., 2001; Yu et al., 2006). Piepho et al. (2008) recapitulates the underlying quantitative genetic assumptions of incorporating a coancestry-based covariance matrix in these models, such as gametic-phase equilibrium of the base population and absence of epistasis, selection and drift. These assumptions are rarely or never honoured in a plant breeding context and even explicitly violated when the genotypes under study represent a set of highly selected inbred lines. However, in practice, despite the numerous deviations from quantitative genetic theory, the CoC often results in an improved model

This chapter has been redrafted after

Maenhout S., De Baets B. and Haesaert G. (2009). Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theoretical and Applied Genetics*, 118:1181-1192.

Maenhout S., De Baets B. and Haesaert G. (2009). CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. *Bioinformatics*, 25:2753-2754.

fit compared to alternative methods for structuring the covariance between the genetic components of inbred lines.

If detailed pedigree information is available for all genotypes under study, one can calculate the CoC by means of the tabular method (Emik and Terrill, 1949). The founding fathers of this pedigree are assumed to be unrelated and therefore set the reference of a zero CoC. Besides accurate pedigree information, the tabular method assumes an equal contribution of both parents to each offspring. The obtained estimators are therefore only valid when there is no selection or genetic drift in the population at hand. However, if inbred lines are obtained through iterative cycles of inbreeding and selection, by doubling haploids or the single seed descent method, the parental contributions are expected to deviate from their theoretical expectations. Molecular marker information allows to circumvent the assumption-burdened pedigree-based estimator, as the resulting allele identities reflect the unequal parental contributions caused by the breeding process. However, deducing the CoC from allele identities on marker loci results in an upwardly biased estimator, because an alikeness in state (AIS) of alleles in different genotypes does not guarantee a shared inheritance from a common ancestor (Cox et al., 1985; Lynch, 1988). Bernardo (1993) shows how this bias can be reduced by taking into account the observed marker similarities between unrelated inbred lines. An alternative approach consists of using marker-based estimation procedures from population genetics, like the kinship coefficient of Loiselle and Graham (1995) or the maximum likelihood estimator described by Thompson (1975), to name but a few. These estimators have their foundations in population genetics but since none of the initial assumptions are met when the genotypes at hand are selected inbred lines, they reduce to the same level as Bernardo's ad hoc method.

Irrespective of the estimation procedure used, the resulting pairwise CoC values are often arranged in a symmetric relationship matrix $\boldsymbol{A}$ which is then used to model the covariance structure between specific components involved in a linear mixed model analysis of genetic evaluation data. This matrix should therefore be at least positive semi-definite (PSD) which implies that all eigenvalues of the matrix are greater than or equal to zero, or equivalently that

$$\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v} \geq 0\,, \qquad \forall \boldsymbol{v} \neq \boldsymbol{0}\,.$$

If we were to model the variance of a vector of random additive genetic effects $\boldsymbol{u}$ as $2\sigma^2_{\text{gca}}\boldsymbol{A}$ (Lynch and Walsh, 1998), $\boldsymbol{A}$ would have to be PSD, as the variance of any linear combination of the additive effects $\text{Var}(\boldsymbol{v}'\boldsymbol{u}) = 2\sigma^2_{\text{gca}}\boldsymbol{v}'\boldsymbol{A}\boldsymbol{v}$ must be positive or 0. A marker-based CoC estimation procedure should therefore guarantee that any derived relationship

matrix is PSD. Unfortunately, most published estimation procedures can result in a non-PSD $\boldsymbol{A}$ matrix, while for those who seem empirically PSD, a formal proof of this property has not been established. Trying to fit a non-PSD covariance structure in a linear mixed model is however not without consequence. Most linear mixed model packages use the PSD property to decompose the variance matrix of the model by means of a Cholesky decomposition. If the variance matrix of the linear mixed model is not PSD, the linear mixed model package either quits with an error message referring to a problem in the initial likelihood calculation (SAS PROC MIXED, Wombat) or gives a warning message and continues the analysis (ASReml). In the latter case, convergence problems of the REML algorithm are frequently observed and the resulting BLUPs should be interpreted with caution as the estimation procedure can now force certain BLUPs to expand away from zero instead of shrinking them.

Several estimation procedures can possibly result in estimated CoC values that are greater than 1 or smaller than 0. From a sheer model fit perspective, a negative covariance between certain genetic components might be justifiable, but when a biological interpretation of the estimated variance components or BLUPs according to Stuber and Cockerham (1966) is needed, the CoC should be a probability and thus bounded by zero and one. To accommodate an interpretation of the CoC estimator according to its original definition, Bernardo (1993) proposes to truncate the out of bound values at the boundaries of the parameter space. As a consequence, even if the used CoC estimation procedure is proven to always generate a PSD relationship matrix, the possibility of a post-hoc truncation of the out of bound values results in a loss of this mathematical property.

If a non-PSD coancestry matrix should arise for whatever reason, it can always be bent towards the closest PSD matrix. The term matrix bending was first coined by Hayes and Hill (1981) for describing a procedure which shrinks the range of eigenvalues of a matrix involved in selection index calculations. The authors indicate, rather as a side-effect, that this procedure allows to make a non-PSD, genotypic or phenotypic variance matrix PSD. More than 20 years later, Sørensen et al. (2002) used this procedure for bending estimated CoC matrices and compared its performance to two other procedures based on spectral decomposition. Unfortunately, all three described procedures allow to obtain CoC values outside the parameter space. Henshall and Meyer (2002) published two programs which focus on bending non-PSD covariance matrices which might arise in multi-trait genetic evaluations. The iterative matrix bender described by Jorjani et al. (2003) focuses on the same problem and allows to give different weights to each entry in the covariance matrix, depending on its reliability. The described algorithm even allows to incorporate the restrictions specific to correlation matrices but these obviously differ from coancestry

matrices.
The main objective of our research was to develop a new marker-based CoC estimation procedure for specific use in hybrid breeding programmes. This procedure should therefore allow for a mix of heterozygous and inbred genotypes. All pairwise CoC values should be interpretable as a probability and therefore lie in the unit interval $[0, 1]$. Any resulting relationship matrix should be guaranteed to be PSD which avoids the need for any bending procedure. In the next section we derive this new estimation procedure and give a formal proof of its PSD property. In the two following sections we compare its behaviour to other CoC estimation procedures by means of simulations and an application to actual maize breeding data. We conclude this chapter by presenting the results of these calculations and a general discussion.

6.2 Materials and methods

6.2.1 WAIS

A codominant molecular fingerprint of a diploid genotype i can be represented as an integer row vector $\boldsymbol{x}_i$. Each position in this vector represents an allele at a certain locus that is represented in the genotyped breeding pool. The vector position is set to 2 if the matching allele is present at both homologous chromosomes of genotype i, 1 in case the allele is present at only one of the two chromosomes and 0 in case of absence. $\boldsymbol{x}_i$ therefore has length $p = \sum_{k=1}^{l} n_k$ where l is the number of genotyped loci and n_k is the number of distinct alleles observed in the collection of genotypes at locus k. Using these vectors we can calculate f_{ij}^{AIS} between two genotypes i and j as

$$f_{ij}^{\mathrm{AIS}} = \frac{1}{4l}\boldsymbol{x}_i\boldsymbol{x}_j{}' \,. \tag{6.1}$$

If we arrange the row vectors $\boldsymbol{x}_i$ of length p for all m genotyped individuals in an $m \times p$ matrix $\boldsymbol{X}$ we can calculate the symmetric AIS (alikeness in state) matrix as

$$\boldsymbol{A}^{\mathrm{AIS}} = \frac{1}{4l}\boldsymbol{X}\boldsymbol{X}' \,.$$

$\boldsymbol{A}^{\mathrm{AIS}}$ can be shown to be at least positive semi-definite (Gower, 1971) as

$$\boldsymbol{v}'\boldsymbol{A}^{\mathrm{AIS}}\boldsymbol{v} = \frac{1}{4l}\boldsymbol{v}'\boldsymbol{X}\boldsymbol{X}'\boldsymbol{v} = \frac{1}{4l}(\boldsymbol{X}'\boldsymbol{v})'(\boldsymbol{X}'\boldsymbol{v}) = \frac{1}{4l}\sum_{z=1}^{p} u_z^2 \geq 0 \,,$$

for all m-sized vectors $\boldsymbol{v} \neq \mathbf{0}$ where $(u_1, u_2, \ldots, u_p)$ is the transpose of the column vector $\boldsymbol{X}'\boldsymbol{v}$.

Despite being PSD, AIS is upwardly biased and therefore not the preferred similarity measure for linear mixed modelling of breeding data. Therefore, we want to incorporate a correction factor without loosing the PSD property. To calculate this correction factor, we start from a normal hybrid breeding scenario which assumes that the inbred lines, for which we want to estimate the pairwise relationships, all belong to the same heterotic group. We also assume that we have a complementary heterotic group of genotyped inbred lines at our disposal. All inbred lines from the first heterotic group are assumed to be completely unrelated to the lines belonging to the second heterotic group. We are now able to define several probabilities that are needed to introduce the correction factor. Imagine we draw a random allele from individuals i and j, at the same locus and both alleles α_i and α_j turn out to be allele z. We define the conditional probability ω_z for two random individuals as

$$\begin{aligned} \omega_z &= \mathrm{P}(\alpha_i \overset{\text{ibd}}{=} \alpha_j \mid \alpha_i = z, \alpha_j = z) \\ &= \frac{\mathrm{P}(\alpha_i = z, \alpha_j = z) - \mathrm{P}(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\cancel{\text{ibd}}}{=} \alpha_j)}{\mathrm{P}(\alpha_i = z, \alpha_j = z)}, \end{aligned} \tag{6.2}$$

where $\mathrm{P}(\alpha_i = z, \alpha_j = z)$ is the probability that the two alleles, drawn from two random individuals i and j of the same heterotic group at the locus to which z belongs, are equal to z and therefore AIS. $\mathrm{P}(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\cancel{\text{ibd}}}{=} \alpha_j)$ is the same probability but with the additional constraint that the alikeness in state is not caused by a shared inheritance from a nearby ancestor (i.e. an ancestor that is still unrelated to all lines in the complementary heterotic group). $\mathrm{P}(\alpha_i = z, \alpha_j = z)$ can be estimated from the $\frac{m(m-1)}{2}$ possible pairs of genotyped members of the heterotic group as

$$\mathrm{P}(\alpha_i = z, \alpha_j = z) = \frac{\sum_{i=1}^{m} \sum_{j>i}^{m} x_{(i,z)} x_{(j,z)}}{2m(m-1)}, \tag{6.3}$$

where $x_{(i,z)}$ and $x_{(j,z)}$ represent the corresponding entries in matrix $\boldsymbol{X}$ for genotypes i and j and the column corresponding to allele z. If we now assume that individual i belongs to one heterotic group and j to another, we can estimate the probability of obtaining an alikeness in state for allele z that did not originate from a shared inheritance from a nearby ancestor. If we define m_1 and m_2 as the number of genotyped members in the first and second heterotic group, respectively, then we can estimate $\mathrm{P}(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\cancel{\text{ibd}}}{=} \alpha_j)$ for both groups as

$$\mathrm{P}(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\not{\mathrm{ibd}}}{=} \alpha_j) = \frac{\sum_{i=1}^{m_1} \sum_{j=m_1+1}^{m_1+m_2} x_{(i,z)} x_{(j,z)}}{4 m_1 m_2}\,, \tag{6.4}$$

where i and j now index over individuals from the first and second heterotic group, respectively. Due to small sample size effects, it is possible that the estimator for $\mathrm{P}(\alpha_i = z, \alpha_j = z, \alpha_i \overset{\not{\mathrm{ibd}}}{=} \alpha_j) > \mathrm{P}(\alpha_i = z, \alpha_j = z)$ in which case the conditional probability ω_z should be set to 0. For rare alleles $\mathrm{P}(\alpha_i = z, \alpha_j = z)$ might be 0 but in those cases the conditional probability is not needed for the calculation of the coancestry. If we now arrange the conditional probabilities ω_z from Eq. (6.2) for each allele z on the diagonal of an all zero square matrix $\boldsymbol{W}$ of size p, we can calculate f_{ij}^{WAIS} for two individuals i and j belonging to the same heterotic group as

$$f_{ij}^{\mathrm{WAIS}} = \frac{1}{4l} \boldsymbol{x}_i \boldsymbol{W} \boldsymbol{x}_j{}'\,, \tag{6.5}$$

where the index WAIS is shorthand for weighted alikeness in state. The procedure thus far has assumed that i and j are different genotypes belonging to the same heterotic group. For the calculation of the symmetric matrix $\boldsymbol{A}^{\mathrm{WAIS}}$ we also need to calculate f_{ii}^{WAIS} for each of the m individuals in the heterotic group. In this case, the conditional probability of Eq. (6.2) underestimates the actual IBD probability and this is even more the case when genotype i has been inbred for g_i generations as is common in hybrid breeding. If we draw two alleles α_{i1} and α_{i2} at the same locus of inbred line i, the conditional probability of Eq. (6.2) should be corrected to

$$\begin{aligned} y'_{i,z} &= \mathrm{P}(\alpha_{i1} \overset{\mathrm{ibd}}{=} \alpha_{i2} \mid \alpha_{i1} = z, \alpha_{i2} = z) \\ &= \frac{1}{2} + \frac{1}{2}\left[1 - \left(\frac{1}{2}\right)^{g_i} + \left(\frac{1}{2}\right)^{g_i} \omega_z\right] \\ &= \left[1 - \left(\frac{1}{2}\right)^{(g_i+1)}\right] + \omega_z \left(\frac{1}{2}\right)^{(g_i+1)}\,, \end{aligned} \tag{6.6}$$

where ω_z is the entry in the diagonal of $\boldsymbol{W}$ corresponding to allele z. If we define

$$\begin{aligned} y_{i,z} &= y'_{i,z} - \omega_z \\ &= \left[1 - \left(\frac{1}{2}\right)^{(g_i+1)}\right](1 - \omega_z)\,, \end{aligned}$$

we can see that $y_{i,z}$ can never be negative. In case all genotyped individuals i in the heterotic group have the same level of inbreeding g we can drop the index i in this last

equation and use the same value y_z for all individuals. For each of the m genotyped individuals in the heterotic group we calculate

$$q_i = \sum_{z=1}^{p} x_{(i,z)}^2 y_{i,z} \,,$$

and arrange these values on the diagonal of an all zero square matrix $\boldsymbol{Q}$ of size m. We can now calculate the WAIS coancestry matrix as

$$\boldsymbol{A}^{\mathrm{WAIS}} = \frac{1}{4l} \left(\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}' + \boldsymbol{Q} \right) . \tag{6.7}$$

The estimated matrix $\boldsymbol{A}^{\mathrm{WAIS}}$ is guaranteed to be PSD as the sum of two PSD matrices $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'$ and $\boldsymbol{Q}$ is always PSD. It is easy to show that $\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'$ is PSD as for any m-sized vector $\boldsymbol{v}$

$$\boldsymbol{v}'\boldsymbol{X}\boldsymbol{W}\boldsymbol{X}'\boldsymbol{v} = (\boldsymbol{X}'\boldsymbol{v})'\boldsymbol{W}(\boldsymbol{X}'\boldsymbol{v}) = \sum_{z=1}^{p} u_z^2 \omega_z \geq 0 \,,$$

where the last inequality follows from the fact that for all alleles z, ω_z is always greater than or equal to zero. Also matrix $\boldsymbol{Q}$ is PSD as it is a diagonal matrix and all entries q_i are greater or equal to zero.

6.2.2 Simulations

In population genetics, the statistical behaviour of marker-based coancestry estimators is usually determined by repeatedly simulating pairs of genotypes for which the true relatedness belongs to a discrete number of predefined classes (Ritland, 1996; Lynch and Ritland, 1999; Van de Casteele et al., 2001; Milligan, 2002). The mean, standard error, bias and possibly other statistical features are examined with loci number, allele number and allele frequency distributions as variables. All of the previously mentioned studies focus on natural populations and therefore assume linkage equilibrium throughout the genome. However, Stich et al. (2005) show the presence of significant linkage disequilibrium (LD) between SSR marker loci of elite, European and US maize germplasm. In a later study, Stich et al. (2007) demonstrate, by means of simulation studies, that selection and drift are the major forces generating this LD. As we want to study the behaviour of different relatedness estimators under realistic breeding circumstances, we must incorporate LD between marker loci. Therefore, each simulation tracks selection by means of several breeding cycles from open-pollinated varieties (OPV) towards elite inbred lines.

The simulations used in this study follow the approach of Stich et al. (2007) and therefore indirectly mimic the breeding scheme of the University of Hohenheim. We assume that the inbred lines are genotyped with 101 microsatellite loci, which are evenly distributed over the maize genome according to a proprietary linkage map of the breeding company RAGT R2n. We also generate 250 QTL loci of the selection trait (e.g. yield) which are randomly positioned on the genetic map. The QTL effects and resulting phenotypic values for line per se and testcross performance were calculated according to Stich et al. (2007). An important difference in the presented simulations is the determination of the number of alleles and the allele frequency distribution of all loci on the map. Stich et al. (2007) use SSR allele frequencies obtained from five Central European OPVs and copy these on the simulated QTLs. Other studies assume identical allele frequency distributions across loci (Ritland, 1996; Lynch and Ritland, 1999) or allow independent draws from a Dirichlet distribution for each locus (Milligan, 2002). We follow the latter approach but also allow the number of alleles to differ between loci. We obtain the number of alleles for each locus as an independent draw from a Poisson distribution plus two, where the parameter λ varies between 0 and 12. This last upper bound was determined by observing little change in the behaviour of the different CoC estimators at higher values of λ.

Each simulation starts by generating an initial base population in Hardy-Weinberg equilibrium. Allele frequencies of each locus are drawn from a Dirichlet distribution with all parameters set to one. From this base generation, we generate the allele frequencies of two subpopulations which have diverged because of artificial selection or geographical differentiation. We assume that on average individuals within each subpopulation share more ancestry compared to individuals belonging to different subpopulations. Wright's Fst value (Wright, 1943, 1951) is a measure for this population stratification and we assume this value to be constant over all loci. The allele frequencies in the subpopulations for locus k are drawn from a Dirichlet distribution with parameters $\theta \boldsymbol{p}_k$ where $\boldsymbol{p}_k$ is the vector of allele frequencies at locus k in the base population and Fst $= \frac{1}{1+\theta}$ (Balding, 2003). 50 individuals are randomly drawn from each of the two populations as an entry point for the first breeding cycle. Each breeding cycle consists of 6 generations of inbreeding and subsequent phenotypical selection based on line per se or testcross performance as described by Stich et al. (2007). This results in 28 almost homozygous inbred lines within each heterotic group (former subpopulation) which are either intercrossed to produce 50 new genotypes for the next breeding cycle or used to compare the different relatedness estimators. For each allele in the breeding pool we keep track of the original founder allele from which it originated. This allows us to calculate the true pairwise CoC values between pairs of inbred lines as an average of the actual IBD relationships over all genotyped loci.

We also calculate the pedigree-based coefficient of coancestry (PED), AIS, WAIS and the estimators described by Bernardo (1993) (BNO), Thompson (1975) (MLE) and Loiselle and Graham (1995) (LOI). Some BNO values are negative while LOI admits to values smaller than 0 or greater than 1. These values are consequently truncated to either 0 or 1 to obtain estimators within the biologically meaningful parameter space.

6.2.3 Maize breeding data

Besides simulations, we use the described relatedness estimators to determine the CoC of the 197 selected inbred lines that are part of the RAGT R2n breeding programme as described in Chapter 4. We slightly adjust one of the initial selection criteria by defining a connected set of METs as a set for which each MET has at least one hybrid or check variety in common with another MET in the selection. As a result, the number of hybrids used in this study is 2367 instead of the original selection of 2361. By contrast, the actual number of METs is reduced from 1284 to 1280 by adding a new selection constraint that requires each MET to test at least three or more hybrids. Fitting an appropriate linear mixed model to the resulting phenotypic data turned out to be problematic due to excessive memory requirements for setting up and solving the mixed model equations. Therefore, the initial selection of 33991 check varieties was reduced to 3022. Only checks that actually connect two or more METs are retained altough METs having less than four checks are filled up with randomly chosen, non-connecting check varieties. We consider all environmental factors as fixed, while the genotypical components and G×E interactions are considered as random effects. The full model for the mean of the vector of phenotypical measurements $\boldsymbol{y}$ can be represented as

$$\begin{aligned} E[\boldsymbol{y}] = \mu &+ \boldsymbol{X}_{(g)}\boldsymbol{g} + \boldsymbol{X}_{(l)}\boldsymbol{l} + \boldsymbol{X}_{(g.l)}\boldsymbol{g.l} + \boldsymbol{X}_{(m)}\boldsymbol{m} \\ &+ \boldsymbol{X}_{(m.l)}\boldsymbol{m.l} + \boldsymbol{X}_{(m.l.t)}\boldsymbol{m.l.t} + \boldsymbol{X}_{(m.l.t.b)}\boldsymbol{m.l.t.b}\,. \end{aligned} \tag{6.8}$$

Here μ represents the global phenotypical mean, while $\boldsymbol{g}$, $\boldsymbol{l}$, $\boldsymbol{m}$, $\boldsymbol{t}$, $\boldsymbol{b}$ represent vectors containing the effects for growing seasons, locations, METs, trials and blocks respectively. The interaction terms in the model are represented as a listing of the appropriate vector symbols, separated by a dot. The $\boldsymbol{X}_{(*)}$ matrices link the effects in each vector to the phenotypical measurements in vector $\boldsymbol{y}$. The effects in vector $\boldsymbol{m}$ are nested within growing seasons, but the METs have received a unique identifier and therefore the notation $\boldsymbol{g.m}$ has been replaced by $\boldsymbol{m}$. We were not able to fit model terms containing treatment effects as we have no information about the specific treatment (irrigation, fertilisation, ...) applied

in each trial. The main effects for year and location are removed from the model for the mean as all their levels are confounded with those of higher level interaction terms. The term $\boldsymbol{X}_{(m.l)}\boldsymbol{m.l}$ was also dropped from Eq. (6.8) as 98% of the location/growing season combinations contain only trials belonging to separate METs. Most of the effects in vector $\boldsymbol{m.l}$ are therefore confounded with the effects in the higher interaction term $\boldsymbol{m.l.t}$. Furthermore, the data contains little or no information for the remaining effects in $\boldsymbol{m.l}$, as different treatments were applied in the few cases where two trials of the same MET were placed within the same location/growing season combination.
The main effects of the random part of the mixed model can be represented as

$$\boldsymbol{Z}_{(c)}\boldsymbol{c} + \boldsymbol{Z}_{(s)}\boldsymbol{a}_s + \boldsymbol{Z}_{(o)}\boldsymbol{a}_o + \boldsymbol{Z}_{(d)}\boldsymbol{d} + \boldsymbol{e}\,. \tag{6.9}$$

Vector $\boldsymbol{c}$ contains the total genotypical effects for all checks, and $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$ are vectors containing GCA effects for the inbred lines belonging to the ISSS and Iodent heterotic groups, respectively. Vector $\boldsymbol{d}$ contains the SCA effects for each of the 2367 hybrids and $\boldsymbol{e}$ contains a residual error for each phenotypical measurement in $\boldsymbol{y}$. The rows of the matrix $\boldsymbol{Z}_{(c)}$ corresponding to measurements on genotyped hybrids are set to 0, while all rows of the remaining $\boldsymbol{Z}$-matrices are set to 0 when their corresponding entries in vector $\boldsymbol{y}$ pertain to check varieties.
Random G×E interaction terms are introduced in the full model for the variance by pairwise interacting the first four model terms in Eq. (6.9) with all the model terms in Eq. (6.8) except $\boldsymbol{m.l}$ and $\boldsymbol{m.l.t.b}$. Due to a software restriction in the maximum number of unknown variance parameters and the prohibitively large computer memory requirements, the possibly improved model fit of factor analytic and reduced rank variance structures for the G×E interaction terms can not be verified. For the same reasons, heterogeneous residual variances can not be fitted. Akaike's information criterion is used to identify the important variance components. At this stage, AIS is used to model the covariance between the general and specific combining abilities of the hybrids according to Stuber and Cockerham (1966). The other random effects are assumed to have a diagonal variance matrix. The described model selection procedure is repeated for the traits grain yield (q/ha at 15% moisture), grain moisture content and days until flowering. The logit transformation is applied to the measurements of grain moisture content as to reduce the skewness in the distribution of the residuals. To avoid convergence problems during REML iterations, these transformed measurements are multiplied with a scaling factor of 100. For both yield and grain moisture content, Akaike's information criterion indicates that the full model for the variance, containing 25 variance parameters, is to be preferred. For days until flowering

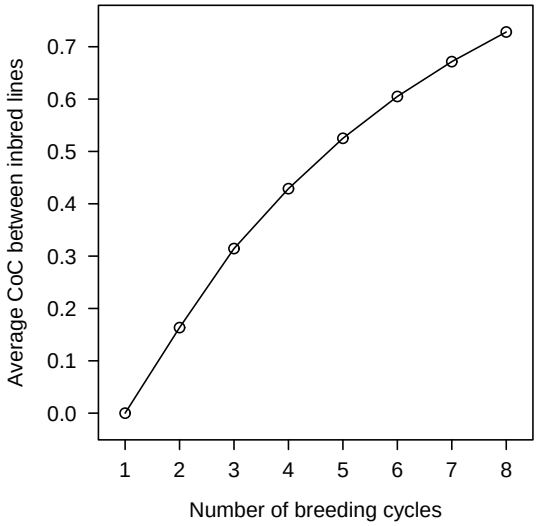
on the other hand, the three interactions between the SCA effects and $\boldsymbol{l}$, $\boldsymbol{m}$ and $\boldsymbol{m.l.t}$ are dropped from the model for the variance. This reduces the number of variance parameters for this trait to 22.

All variance components are estimated through REML optimisation by means of the Average Information algorithm as implemented in the software tool ASReml (Gilmour et al., 2002). The model fit of the different CoC matrices, obtained by applying each of the examined procedures, is determined by replacing them for the AIS-based matrices in the covariance models of the vectors $\boldsymbol{a}_s$, $\boldsymbol{a}_o$ and $\boldsymbol{d}$ in Eq. (6.9) and evaluating the resulting restricted log-likelihood at the end of the REML iteration. Both BNO and LOI produce CoC values that are outside the biologically meaningful parameter space ($0 \leq f_{ij} \leq 1$ for all genotypes i and j) and these values are therefore truncated at the boundaries. For both heterotic groups the MLE and the bounded LOI numerator matrices are non-PSD and therefore need bending towards the nearest PSD matrix.

6.2.4 Bending procedures

The first examined bending procedure applies a spectral decomposition of the non-PSD matrix and replaces all negative eigenvalues with a small positive value as described in Sørensen et al. (2002) and Jorjani et al. (2003). This procedure does however not constrain the elements of the bended matrix within the unit interval such that new boundary infringements might arise during bending. To enforce these boundary constraints we implemented an MCMC procedure, inspired by FLBEND (Henshall and Meyer, 2002), to transform non-PSD coancestry matrices towards the closest PSD matrix within the parameter space. The idea behind FLBEND is to generate a symmetric matrix $\boldsymbol{B}$ by means of an iterative Monte Carlo procedure such that the distance between the PSD matrix product $\boldsymbol{BB'}$ and the non-PSD input matrix $\boldsymbol{A}$ is minimised. Perturbations in $\boldsymbol{B}$ that increase this distance are accepted at reduced probability. Our modified algorithm rejects alterations in $\boldsymbol{B}$ that allow the elements of $\boldsymbol{BB'}$ to stray outside the unit interval. To allow for a faster convergence under this restricted setting, we continuously update the variance of new perturbations by means of a Metropolis-Hastings step. We also allow the matrix $\boldsymbol{B}$ to be non-symmetrical as this results in a better approximation of the input matrix, at the cost of a higher computational demand. The pseudo-code for this MCMC bending procedure is given in the appendix of this chapter.

Figure 6.1: Average coefficient of coancestry between inbred lines at each breeding cycle



6.3 Results

6.3.1 Simulated breeding populations

The first breeding cycle in each simulation produces 28 unrelated inbred lines. The selective pairwise mating of these inbred lines produces 50 hybrids, which represent the starting point for the next selection cycle. At the end of each breeding cycle we can calculate the actual CoC between all pairs of inbred lines by averaging over the true IBD relationships at the SSR marker loci. Figure 6.1 depicts this average CoC at each breeding cycle.
At the end of each breeding cycle, only the best performing inbred lines are retained, regardless of their pairwise relatedness. This behaviour mimics a real hybrid breeding programme where decisions are based on phenotypical performance data. Unfortunately this implies that it is not possible to control the pairwise CoC between inbred lines at each breeding cycle which would allow to quantify the standard error of the different estimators at predefined levels of coancestry (parent-offspring, half-sibs,...). Instead, we determine for the 378 pairwise combinations of the 28 selected inbred lines within a heterotic group the actual CoC based on the average IBD relationships at the SSR loci. This allows us to

determine the average bias and root mean squared error (RMSE) of each CoC estimator. Figure 6.2 visualises for each estimator this RMSE at different values of the Fst between the initial OPVs and different values of λ. The presented RMSE values are averaged over 100 independent iterations of the simulation routine.
The AIS, WAIS and PED estimators are guaranteed to produce a PSD coancestry matrix, while the other three estimators (BNO, MLE and LOI) are not. For every CoC estimator the proportion of non-PSD numerator relationship matrices was determined by means of an eigenvalue analysis. During simulations, BNO never resulted in a non-PSD $\boldsymbol{A}^{\text{BNO}}$ matrix despite the fact that several small truncations were necessary to confine the estimator within the biologically meaningful parameter space. MLE and LOI are more likely to produce a non-PSD coancestry matrix as can be seen from Figure 6.3.

6.3.2 Maize breeding data

AIS, PED, MLE and WAIS all produce CoC estimators within the unit interval. BNO on the other hand can produce negative CoC values and the same holds for LOI which also allows to obtain CoC values greater than one. Figure 6.4 shows the range of pairwise CoC values between genotypes belonging to the same heterotic group for all examined estimation procedures.
The covariance structure of the GCA and SCA effects of Eq. (6.9) is modelled by means of the six examined coancestry estimators. We can compare the goodness-of-fit of these coancestry estimators by means of the restricted log-likelihood at the final REML iteration, as the fixed effects structure and the number of estimated variance components for each model are constant. To allow for a fair comparison between estimators we restrict all CoC values to lie within the unit interval. This decision only has a minor effect on the model fit as the difference in restricted log-likelihoods between the bounded and unbounded variants of BNO and LOI is negligible for all three traits under study. MLE is bounded by nature but results in non-PSD CoC matrices for both the Iodent and the ISSS heterotic groups and so does the bounded LOI variant. The MCMC bending procedure results in a smaller distance between the original non-PSD matrix and the bended output matrix compared to the spectral decomposition approach. For the MCMC bending procedure the maximum element-wise average distance is only 0.00075, while it is 0.0015 for the spectral decomposition approach. This superiority is however barely reflected in an improved model fit as the restricted log-likelihoods of the LOI and MLE CoC matrices bended with the MCMC procedure are usually identical or slightly higher than those bended with the spectral de-

Figure 6.2: Root mean squared error of each CoC estimator at the different stages of a hybrid breeding programme. Panels are sorted according to the Fst value of the two initial OPVs from which the selection routine started and λ, the expected value of the Poisson distribution which was used to draw the number of alleles at each locus. RMSE values are averaged over 100 iterations of the simulation routine.

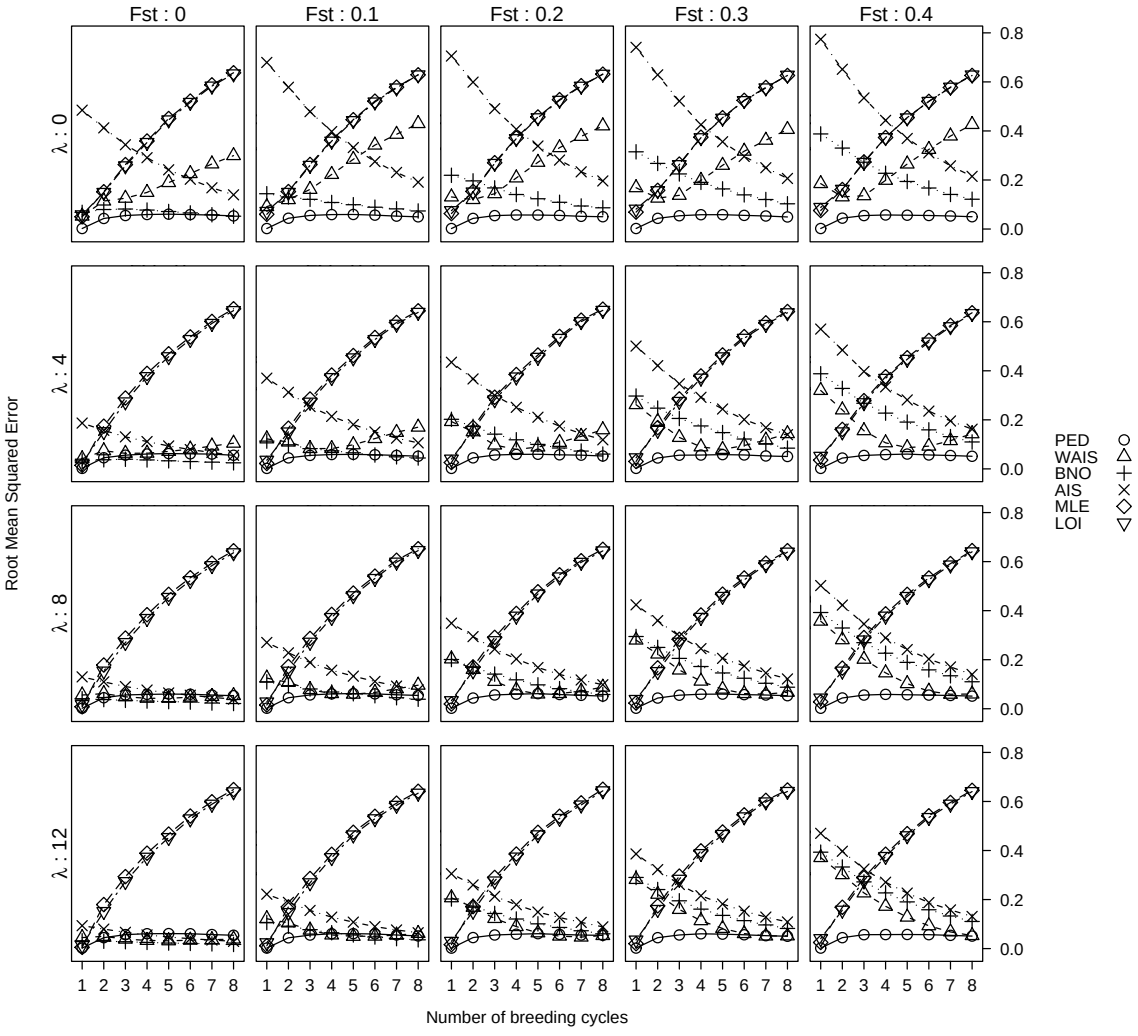


Figure 6.3: Proportion of non-PSD coancestry matrices for MLE and LOI

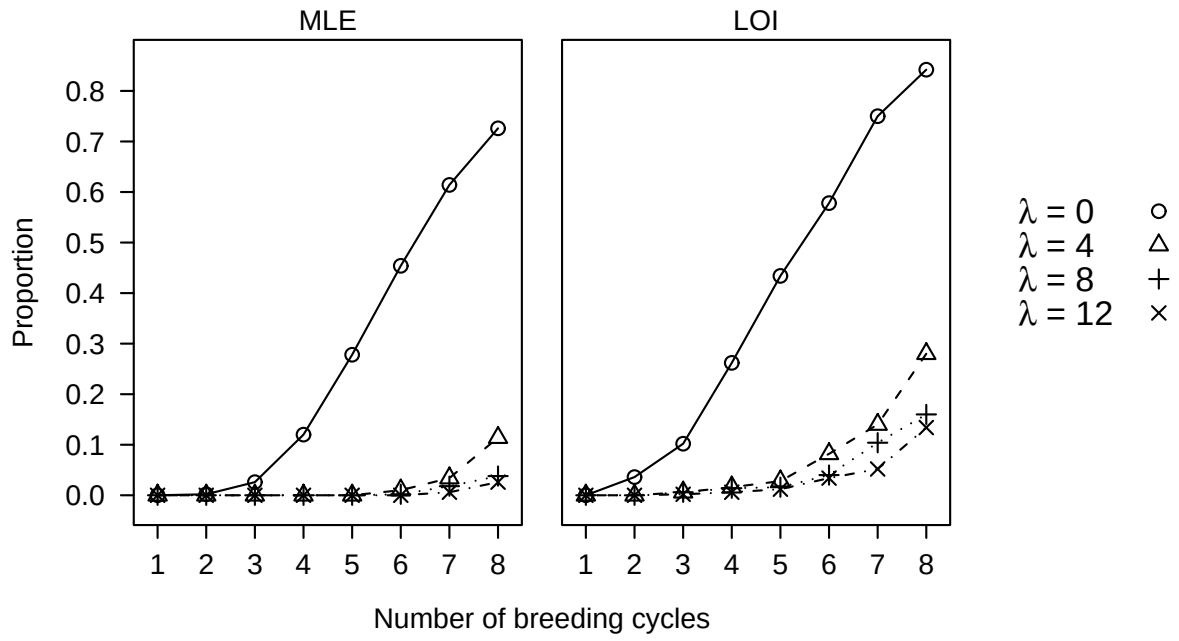


Figure 6.4: Pairwise CoC values between members of the same heterotic group, estimated by means of the 6 examined procedures. For LOI and BNO both the unbounded (suffix u) and bounded (suffix b) ranges are presented.

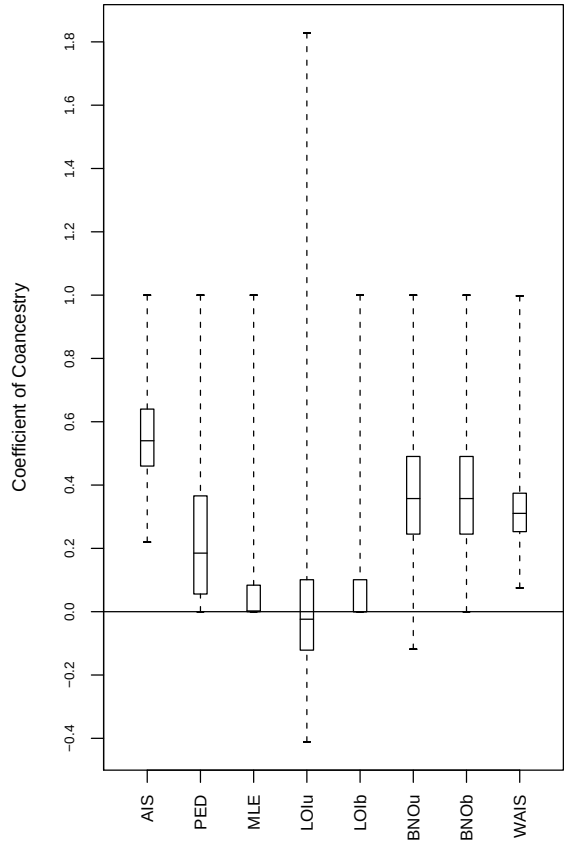


Table 6.1: Restricted log-likelihoods for each of the six coancestry estimators that were used to model the covariance for GCA and SCA effects in Eq. (6.9) for the traits yield, grain moisture content and days until flowering. The number between brackets represents the relative ordering of the estimators when sorted according to decreasing restricted likelihood. BNO and LOI values were bounded within the unit interval. MLE and the bounded LOI matrices were bended towards the closest PSD matrix using the MCMC algorithm.

	yield	moisture %	flowering
PED	−222740.1 (4)	−194696.6 (1)	−55339.6 (1)
AIS	−222734.8 (2)	−194710.8 (2)	−55343.8 (2)
BNO	−222734.8 (1)	−194712.9 (3)	−55344.1 (3)
WAIS	−222739.2 (3)	−194715.3 (4)	−55347.7 (4)
MLE	−222743.2 (6)	−194716.2 (5)	−55357.0 (5)
LOI	−222741.0 (5)	−194725.6 (6)	−55361.9 (6)

composition approach. Table 6.1 gives an overview of the restricted log-likelihoods for each of the examined CoC estimators and for each of the three traits under study.

6.4 Discussion

The CoC is often used to model the covariance between genetic components of genotypes under selection, despite the inherent conflicts with the underlying quantitative genetic theory. In hybrid breeding programmes and certain association studies the genotypes at hand are highly selected inbred lines with little or no information concerning their selection history. Analysing phenotypical data originating from such inbred lines or their pairwise matings by means of a linear mixed model which uses a CoC estimator to model the covariance between GCA or SCA components, should be considered as an approximation, since the resulting variance components and BLUPs are biased. Nevertheless, good results have been obtained in practice using different CoC estimators based on pedigree or molecular marker information.

In this chapter we present a PSD, codominant marker-based relatedness estimator called the weighted alikeness in state or WAIS estimator. This estimator is only applicable in

the specific case that a reference set of genotyped individuals, unrelated to the genotypes in the sample, is available. As hybrid breeders make extensive use of unrelated heterotic groups, this estimator is particularly suited for this type of selection. It should be clear that WAIS is not claimed to be assumption free as it for example relies on gametic phase equilibrium. This assumption is surely not met in advanced breeding pools, so we study the behaviour of WAIS and other CoC estimators under a typical hybrid breeding selection scheme by means of simulations and actual breeding data.

6.4.1 Marker-based estimators

Bernardo (1993) uses the observed marker similarities between unrelated lines to correct the AIS-based estimator for lines belonging to the same heterotic group. Besides the often violated assumption of gametic-phase equilibrium between loci, there is also the problem of obtaining negative values for BNO when the correction factor exceeds AIS.
Thompson (1975) demonstrates how the pairwise relationship between non-inbred individuals can be estimated by means of a likelihood function that incorporates the three possible identity by descent probabilities (Jacquard, 1974). Milligan (2002) compares the behaviour of this estimator to 5 prominent, non-likelihood estimators (Queller et al., 1989; Li et al., 1993; Ritland, 1996; Lynch and Ritland, 1999; Wang, 2002). He concludes that under all simulated scenarios, the maximum likelihood estimator (MLE) exhibits a lower variation compared to the other estimators. However, this reduction in standard error comes at a price, as the likelihood estimator shows considerably more bias, especially at the boundary of the parameter space. A second advantage lies in the fact that the likelihood maximisation procedure is constrained to produce biologically meaningful results ($0 \leq \text{MLE} \leq 1$), but this property could in fact be enforced on the other estimators as well, again at the cost of increasing the bias. Nevertheless, we consider the maximum likelihood estimator to be the most appropriate candidate for use in breeding pools as it explicitly handles inbred individuals. Other implicit assumptions like linkage equilibrium between marker loci and exact knowledge of population allele frequencies are most likely to be violated when the fingerprinted genotypes are all inbred lines but this is the case for all other estimators as well. Anderson and Weir (2007) extended the maximum likelihood approach for the case where the examined genotypes belong to subpopulations of a population with known allele frequencies. However, we did not adopt this approach as its resulting coancestry measures refer to the ancestral population, while all other examined estimators refer to the subpopulation itself.
The problem of finding the most likely IBD relationship between two genotypes can be

formulated as the maximisation of a function over a vector $\boldsymbol{\Delta}$ containing 9 single-locus, identity by descent modes (Jacquard, 1974; Thompson, 1975). As the simulations in Milligan (2002) assume large, non-inbred populations, the parameter space can be reduced to having 2 dimensions. Hepler (2005) explores the possibility of inbred individuals which expands the parameter space to 8 dimensions. Both Milligan (2002) and Hepler (2005) use the downhill simplex method (Nelder and Mead, 1965), a heuristic optimisation technique, because an algebraic solution of the maximisation problem is not feasible. The original version of this heuristic neither allows the incorporation of the boundary constraints ($0 \leq \Delta_i \leq 1$) nor the linear constraint ($\sum_{i=1}^{9} \Delta_i = 1$). Hepler (2005) introduces these constraints by rejecting solutions outside the parameter space during the optimisation process. This results in numerous lost iterations, especially when certain values of $\boldsymbol{\Delta}$ are near the boundary of the parameter space. To allow for simulations to be performed in an acceptable time frame, we use a quasi-Newton nonlinear interior-point method (Meza et al., 2007) for the maximisation of $L(\boldsymbol{\Delta})$. This approach reduces the needed processor time per genotype pair drastically, while the resulting estimators of $\boldsymbol{\Delta}$ are always nearly identical compared to those of the constrained simplex algorithm. The resulting matrix $\boldsymbol{A}^{\mathrm{MLE}}$, containing all pairwise estimates of MLE, is not guaranteed to be PSD which limits its use in a mixed model setting. If the $\boldsymbol{A}^{\mathrm{MLE}}$ happens to be non-PSD, the nearest PSD matrix should be used instead.

Loiselle and Graham (1995) describe a marker-based coancestry estimator which quantifies the correlation in allele frequencies between two individuals belonging to a population in Hardy-Weinberg equilibrium. Despite the obvious violations of underlying theoretical assumptions, this marker-based estimator is sometimes used to model the covariance between genotypes originating from breeding programmes (Yu et al., 2006; Zhang et al., 2007; Casa et al., 2008). LOI is not guaranteed to lie within the parameter space so truncations are often necessary at the boundaries. The resulting coancestry matrix $\boldsymbol{A}^{\mathrm{LOI}}$ is not guaranteed to be PSD.

6.4.2 Simulations

The simulated selection scheme follows the maize breeding programme of the University of Hohenheim (Stich et al., 2007). Each breeding cycle consists of 6 generations of inbreeding and selection after which the best performing inbred lines are mated to provide the initial population for the next breeding cycle. The actual CoC, obtained as an average over SSR loci, gradually increases as the number of subsequent breeding cycles rises. The trend observed in Figure 6.1 is independent from the initial parameter selection (Fst, λ), which

indicates that the number of breeding cycles can be used as an indirect measure for the average relatedness within the heterotic groups.

From Figure 6.2 we can see that the pedigree-based estimator outperforms all marker-based estimators by producing the lowest root mean squared error under all parameter settings. The bias introduced by unequal parental contributions as a consequence of the selection process seems to be negligible compared to the bias of the marker-based estimators. The advantage of the pedigree-based estimator might not be so apparent under practical breeding circumstances, as detailed and accurate pedigree records, tracing back to the initial OPVs, are usually not available. Looking at the marker-based estimators we see that the behaviour of MLE and that of LOI are nearly identical under all simulated scenarios. The performance of these estimators deteriorates as the number of breeding cycles increases which is probably caused by the increasing deviations from population genetics assumptions on which they rely. AIS shows a rather reversed picture as it tends to become more accurate as the number of breeding cycles increases. The overestimation of AIS at low levels of selection is more pronounced when the expected number of distinct alleles at each locus (λ) is small or the differentiation between the populations from which the heterotic groups are developed (Fst) is large. The influence of the value of the Fst is rather surprising as AIS makes no use of a reference population. A possible explanation might lie in the constraints that are imposed on the allele frequencies as a consequence of fixing the Fst value. This might have the same effect as lowering the effective number of distinct alleles at each locus.

The RMSE of WAIS and BNO is usually at a considerably lower level compared to the other marker-based estimators. When $\lambda = 0$, which is equivalent to fixing the number of distinct alleles of each SSR or QTL locus at 2, WAIS has a higher RMSE compared to BNO. This rather unrealistic scenario allows AIS to outperform WAIS when the number of breeding cycles is high. As soon as λ increases to a more realistic setting, WAIS outperforms AIS and can compete with BNO. Ho et al. (2005) estimate the Fst between Corn Belt dent populations to be 0.142 which is somewhat similar to the 0.15 found earlier by Labate et al. (2003). At this level of differentiation WAIS and BNO perform at a comparable level, although WAIS performs slightly better when allelic diversity is high. WAIS also outperforms BNO when the Fst value increases, except when λ is small and the number of breeding cycles is high.

WAIS is specifically designed to guarantee a PSD coancestry matrix. This property is necessary when this matrix is used to model the covariance between genetic components in a linear mixed model. BNO, despite not being a PSD estimator, always produced a PSD coancestry matrix for all simulated populations and the real hybrid maize data

set. Several truncations towards 0 were necessary but these were small in absolute value. These arguments allow to conclude that BNO is a stable estimator which produces natural coancestry measures under variable circumstances. This cannot be said for MLE and LOI which both produced non-PSD coancestry matrices for a rather large proportion of the simulated heterotic groups. This proportion is highly dependent on the number of distinct alleles at each locus where a value of λ of 0 and 8 consecutive breeding cycles results in a very high probability of obtaining a non-PSD matrix. LOI performs slightly worse than MLE, but both estimators generally exhibit the same increase in proportion of non-PSD matrices when the allelic diversity decreases.

6.4.3 Maize breeding data

Figure 6.4 shows that BNO and LOI both produce negative CoC estimates and that LOI also allows the estimators to become greater than one. The infractions of BNO on the lower bound are rather limited in frequency as well as in size as only 37 of all 9843 estimates are smaller than zero with a mean negative deviation of 0.04. Truncation of BNO at the lower bound therefore has little impact on the model fit. LOI on the other hand, ranges from -0.41 to 1.83 and only 43 percent of the estimated CoC values fall within the biologically meaningful parameter space. Truncation of LOI at the boundaries therefore cripples the distribution of CoC values as more than half of the estimates are set to 0 or 1. The bounded LOI distribution looks very similar to that of MLE, the other estimator from population genetics. This is to be expected as MLE forces all estimates to lie within the unit interval by means of the constrained optimisation algorithm. The other estimators produce more natural looking distributions where AIS is generally at a higher level than PED and both BNO and WAIS take more intermediate positions. The unbounded BNO and LOI result in PSD CoC matrices for both the ISSS and Iodent heterotic groups while MLE produces non-PSD matrices. After bounding of BNO and LOI only LOI results in non-PSD CoC matrices for both heterotic groups such that bending needs to be applied.
For the non-PSD matrices produced by MLE and the bounded variant of LOI, two matrix bending procedures were examined. The spectral decomposition approach is computationally quite fast but does not allow to constrain the elements within the unit interval. The application of this bending procedure to the bounded LOI estimator results for example in 2770 new boundary infringements, though it should be noted that these are rather small in absolute value. The MCMC procedure is computationally quite demanding but allows to constrain all CoC values within the aforementioned range and produces a PSD matrix that is closer to the original input matrix than the matrix resulting from the spectral decom-

position approach. This difference between both bending procedures is however negligible when comparing restricted log-likelihoods of linear mixed models in which the bended CoC matrices are used to model covariances between random GCA and SCA effects.
In Table 6.1 we can see that PED results in the highest restricted log-likelihood at the end of the REML iterations for the traits grain moisture content and days until flowering, while BNO, AIS and WAIS outperform the pedigree estimator for the trait yield. If we focus on the marker-based estimators, we see that the uncorrected AIS results in the highest restricted log-likelihood for the traits grain moisture content and days until flowering while for yield the difference with BNO is negligible. Although surprising at first, this behaviour of AIS is consistent with the simulations as it was shown that the RMSE of AIS decreases to that of BNO and WAIS when the number of consecutive breeding cycles is high. Taking into account that AIS always results in a PSD coancestry matrix, this estimator deserves a revaluation when applied to highly selected breeding material. When summing over rank scores, BNO takes third position while WAIS takes fourth. Constraining the resulting coancestry matrix to be PSD comes at the price of a slightly reduced model fit. MLE and LOI, both originating from population genetics, give the lowest log-likelihoods for all three traits under study.
Results from this study indicate that the pedigree-based CoC estimator is superior to the available marker-based alternatives when accurate and complete pedigree information is available for a set of highly selected inbred lines. Comparisons between marker-based CoC estimation procedures, for the specific case that the inbred lines are subdivided in unrelated heterotic groups, indicate that procedures from population genetics like MLE or LOI should generally be avoided as a considerable deviation from the actual IBD relationship can be observed when the inbred lines have a long breeding history. Results also indicate that in this specific case, the observed allele identities need little correction and therefore AIS results in a good approximation of the true CoC. However, if the breeding history is not that long or unknown, BNO and WAIS should be used. BNO generally results in a slightly better model fit, but when the PSD property of the resulting CoC matrix needs to be guaranteed, for example when used in a linear mixed model for breeding value estimation or an association study, the new relatedness estimator WAIS should be preferred.

6.5 CoCoa

The software package CoCoa allows researchers to apply the newly developed WAIS coancestry estimation procedure to their own panels of genotyped plants or animals. Besides WAIS,

CoCoa also implements each of the four examined competitors namely AIS, BNO, LOI and MLE. The molecular fingerprint of each genotyped individual can be provided in the exact same file format as is needed by the software package "structure" (Pritchard et al., 2000). All implemented estimation procedures work exclusively on multilocus genotype data so there is no need to provide pedigree information. Similarly, all estimators assume that the genotyped markers are in gametic phase equilibrium, so genetic map information is not required. This latter assumption, however, implies that all implemented coancestry estimators are most likely biased when applied to genotypes that are part of a breeding program. In fact, under these circumstances, it is impossible to obtain exact, marker-based coancestry estimates and one should try to identify the estimation procedure that introduces the least amount of bias for the set of genotyped individuals at hand.
Besides implementing five different coancestry measures, CoCoa also provides a set of matrix manipulation tools that are useful for dealing with marker-derived covariance matrices. Matrix elements can be bounded within a predefined interval (e.g. the unit interval) and if the estimated matrix is well-conditioned, a matrix inversion routine based on the singular value decomposition allows to obtain a numerically stable inverse for use in linear mixed model packages. The sensitivity to small perturbations, caused for example by round-off errors, is expressed as the 2-norm condition number of the estimated coancestry matrix. The larger this condition number, the less reliable the matrix inverse will be. To decrease the matrix condition number, CoCoa allows to bend the estimated coancestry matrix by means of the two described bending algorithms, based on spectral decomposition and Monte Carlo sampling respectively. CoCoa allows to export the estimated coancestry matrices to various file formats which are required by the most commonly used software packages for linear mixed model analysis such as SAS Proc Mixed, ASReml and Wombat. More general export formats such as a dense rectangular array or lower and upper triangular arrays are also available.
The core components of CoCoa are written in C++, while the graphical user interface is written in Java. CoCoa is published under the terms of the GNU General Public License and can be downloaded free of charge at http://webs.hogent.be/cocoa. Source code, manual, binaries for 32 and 64-bit Linux systems and an installer for Microsoft Windows are provided.

6.6 Appendix: pseudo-code for the MCMC bending algorithm

Function: $\boldsymbol{A}^* \leftarrow BendIt(\boldsymbol{A}, maxD, maxT)$

Input: $\boldsymbol{A}$, a symmetric, non-PSD matrix of size n

Input: $maxD$, allowed distance between input and output matrix

Input: $maxT$, maximum computation time

Output: $\boldsymbol{A}^*$, a symmetric, PSD matrix of size n

$sigma \leftarrow 0.1$; $probsigma \leftarrow minvalue$; $lowestD \leftarrow maxvalue$

$\boldsymbol{B}$, $\boldsymbol{C}$, $\boldsymbol{A}^* \leftarrow DiagonalMatrixOfOnes(n)$

$D \leftarrow DistanceBetweenMatrices(\boldsymbol{A}, \boldsymbol{A}^*)$

while $ElapsedTime < maxT$ **and** $lowestD > maxD$ **do**

- $sigma^* = sigma + RandomGaussian(0, 0.01)$
- $numaccept \leftarrow 0$
- **for** $l \leftarrow 1$ **to** n^2 **do**
  - $accept \leftarrow BendCycle(sigma^*, n, D, \boldsymbol{A}, \boldsymbol{A}^*, \boldsymbol{B})$
  - **if** $accept =$ **true then**
    - $numaccept \leftarrow numaccept + 1$
- $probsigma^* = numaccept/n^2$
- $\alpha = probsigma^*/probsigma$
- $u = RandomUniform(0, 1)$
- **if** $u < \alpha$ **then**
  - $sigma \leftarrow sigma^*$; $probsigma \leftarrow probsigma^*$;
- **if** $D < lowestD$ **then**
  - $lowestD \leftarrow D$
  - $\boldsymbol{C} \leftarrow \boldsymbol{B}$
- **else**
  - $\boldsymbol{B} \leftarrow \boldsymbol{C}$
  - $\boldsymbol{A}^* \leftarrow \boldsymbol{B}\boldsymbol{B}'$
  - $D \leftarrow lowestD$

Function: $accept \leftarrow BendCycle(sigma^*,n,D,\boldsymbol{A},\boldsymbol{A}^*,\boldsymbol{B})$

Input: $sigma^*$, standard deviation
Input: n, size of $\boldsymbol{A}$
Input: $\boldsymbol{A}$, symmetric matrix of size n
Input / Output: $\boldsymbol{A}^*$, symmetric matrix of size n
Input / Output: $\boldsymbol{B}$, matrix of size n
Input / Output: D, current distance between $\boldsymbol{A}$ and $\boldsymbol{A}^*$
Output: $accept$, boolean indicating acceptance of change

$legal \leftarrow$ **true**
$accept \leftarrow$ **false**
$r \leftarrow SelectRandomInt(n)$
$c \leftarrow SelectRandomInt(n)$
$newvalue \leftarrow \boldsymbol{B}(r,c) + RandomGaussian(0, sigma^*)$
$newD \leftarrow D$
for $k \leftarrow 1$ **to** n **do**
  if $k = r$ **then**
    $newrow(k) = \boldsymbol{A}^*(r,r) - \boldsymbol{B}(r,c)^2 + newvalue^2$
  else
    $newrow(k) \leftarrow \boldsymbol{A}^*(r,k) + (newvalue - \boldsymbol{B}(r,c)) * \boldsymbol{B}(k,c)$
  if $OutsideParamSpace(newrow(k))$ **then**
    $legal \leftarrow$ **false**
  else
    $newD \leftarrow newD + AbsValue(\boldsymbol{A}(r,k) - newrow(k)) -$
    $AbsValue(\boldsymbol{A}(r,k) - \boldsymbol{A^*}(r,k));$
if $legal =$ **true then**
  $u = RandomUniform(0,1)$
  if $u < distance/newD$ **then**
    $accept =$ **true**
if $accept =$ **true then**
  $D \leftarrow newD$
  $\boldsymbol{B}(r,c) \leftarrow newvalue$
  for $k = 1$ **to** n **do**
    $\boldsymbol{A}^*(r,k) = \boldsymbol{A}^*(k,r) = newrow(k)$

CHAPTER 7

Support vector machine regression for the prediction of hybrid maize performance

7.1 Introduction

For several agronomically important plant species like maize (*Zea mays* L.), hybrid varieties constitute a considerable part, if not all, of the commercial market. Maize breeding programmes typically have a continuously evolving breeding pool at their disposal which is loosely divided into several complementary heterotic groups. New inbred lines are created by subsequent inbreeding of an initial cross or the use of doubled haploids. During their selection, these candidate lines are crossed with tester lines from a complementary heterotic group and hybrid performance is evaluated in multi-location field trials. Bernardo (1994, 1995, 1996a,c) uses linear mixed modelling to predict the performance of such an untested cross based on field trial results of related hybrids and marker data. This approach performs well considering the upper limit in prediction accuracy that is imposed by the heritability of each tested trait. Charcosset et al. (1998) show that this prediction method is superior when hybrids originate from crosses between unrelated inbred lines, which is most likely the case in commercial breeding programmes. Unfortunately, corre-

This chapter has been redrafted after

Maenhout S., De Baets B., Haesaert G. and Van Bockstaele E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theoretical and Applied Genetics*, 115:1003-1013.

Maenhout S., De Baets B., Haesaert G. and Van Bockstaele E. (2008). Marker-based screening of maize inbred lines using support vector machine regression. *Euphytica*, 161:123-131.

lations between predicted and observed SCA values are too low to allow for an effective selection towards high heterosis hybrids (Bernardo, 1995).
This chapter explores the use of ε-insensitive support vector machine regression (ε-SVR) to predict the phenotypical performance of untested hybrids. The presented technique uses linear mixed modelling to correct unbalanced phenotypical measurements for nuisance parameters like trial, location and block effects. The corrected phenotypical values of all hybrids are used as a training set for constructing an ε-SVR model in which the molecular fingerprints of each hybrid serve as predictor variables. These models can subsequently be used to predict the phenotypical values of unknown hybrids and inbred lines. The advantage of ε-SVR lies in the use of kernel functions that allow to explore non-linear models for hybrid prediction.

7.2 Materials and methods

7.2.1 Data description

The phenotypic data used in this study originate from field trials that were organised as part of the grain maize breeding programme of RAGT R2n as described in Chapter 4. The pedigree backgrounds of the 105 ISSS and 92 Iodent lines were reconstructed as far as possible, which allowed to estimate the pairwise coefficients of coancestry f_{ij}^{PED} between inbred lines belonging to the same heterotic group by means of a tabular analysis with corrections for inbreeding and backcrossing (Emik and Terrill, 1949).
The 197 selected inbred lines were fingerprinted using 101 co-dominant SSR markers. As described in Chapter 4, only 75 SSR markers have complete profiles over all selected inbred lines. In this study only information of these complete SSR loci was used. 2.6% of all SSR locus/inbred line combinations was heterozygous, preventing an exact deduction of the hybrid genotype when these lines are used as parents. The molecular coefficient of coancestry f_{ij}^{BNO} between two inbred lines i and j of the same heterotic group was calculated from their SSR-based fingerprints as (Bernardo, 1993),

$$f_{ij}^{\mathrm{BNO}} = \frac{f_{ij}^{\mathrm{AIS}} - \frac{1}{2}(\bar{f}_{i.}^{\mathrm{AIS}} + \bar{f}_{.j}^{\mathrm{AIS}})}{1 - \frac{1}{2}(\bar{f}_{i.}^{\mathrm{AIS}} + \bar{f}_{.j}^{\mathrm{AIS}})}\,, \tag{7.1}$$

where f_{ij}^{AIS} is the average allele identity over all SSR marker loci between inbred i and j which can be calculated according to Eq. (6.1) or in a more classic notation as

$$f_{ij}^{\mathrm{AIS}} = \frac{1}{4l}\sum_{k=1}^{l} I(i_{k_m}, j_{k_m}) + I(i_{k_m}, j_{k_p}) + I(i_{k_p}, j_{k_m}) + I(i_{k_p}, j_{k_p})\,,$$

where i_{k_m} represents the maternal allele of inbred line i for locus k, while i_{k_p} represents the paternal allele. $I(i_{k_m}, j_{k_m})$ returns 1 if the maternal allele on locus k of individual i is equal to the maternal allele of that same locus of individual j and 0 otherwise. $\bar{P}_{i.}$ represents the average allele identity between inbred line i and all inbred lines of the complementary heterotic group. $\bar{f}_{i.}^{\text{AIS}}$ represents the average allele identity between inbred line i and all lines of its complementary heterotic group.
AFLP fingerprints were generated using 11 *Pst*I-*Mse*I and 4 *Eco*RI-*Mse*I primer combinations. These 15 primer combinations produced 569 polymorphic bands for the 197 selected inbred lines. To calculate the molecular coefficient of coancestry f_{ij}^{JAC} between two inbred lines i and j of the same heterotic group based on dominant AFLP marker data, Eq. (7.1) was used but the proportion of shared AFLP alleles f_{ij}^{AIS} was calculated according to the Jaccard similarity measure as

$$f_{ij}^{\text{AIS}} = \frac{a_{ij}}{a_{ij} + b_{ij} + c_{ij}}\,, \tag{7.2}$$

where a_{ij} represents the number of bands common to both individuals i and j while b_{ij} represents the number of bands unique to i and c_{ij} those unique to j.

7.2.2 Linear mixed model analysis

As the data suffers from severe unbalancedness, a linear mixed model is the recommended approach for correcting the phenotypical measurements for nuisance factors like trial, location and block effects. The used model is quite similar to that proposed by Bernardo (1994) but the actual plot measurements are used instead of averages over locations and blocks:

$$\boldsymbol{y} = \mu + \boldsymbol{X}_{(m)}\boldsymbol{m} + \boldsymbol{X}_{(m.l.t)}\boldsymbol{m.l.t} + \boldsymbol{X}_{(m.l.t.b)}\boldsymbol{m.l.t.b} + \boldsymbol{Z}_{(c)}\boldsymbol{c} + \boldsymbol{Z}_{(s)}\boldsymbol{a}_s + \boldsymbol{Z}_{(o)}\boldsymbol{a}_o + \boldsymbol{Z}_{(d)}\boldsymbol{d} + \boldsymbol{e}\,. \tag{7.3}$$

This formulation follows the notation that was introduced in Chapter 6. $\boldsymbol{y}$ represents a vector containing the trait responses for each plot in the data set and μ represents the global phenotypical mean. $\boldsymbol{m}$ is a vector containing the fixed multi-location trial (MET) effects, nested within growing seasons. $\boldsymbol{m.l.t}$ contains the fixed effects for each trial, nested within a location and nested within a MET. $\boldsymbol{m.l.t.b}$ represents the fixed block effects, nested within each trial. Vector $\boldsymbol{c}$ contains the random genotypical effects for all checks. $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$ are vectors containing GCA effects for the inbred lines belonging to the ISSS and Iodent heterotic groups respectively, while $\boldsymbol{d}$ contains the SCA effects for each of the 2361 hybrids. $\boldsymbol{e}$ contains a random residual error for each plot in the data

set. The coincidence matrices $\boldsymbol{X}_{(m)}$, $\boldsymbol{X}_{(m.l.t)}$, $\boldsymbol{X}_{(m.l.t.b)}$, $\boldsymbol{Z}_{(c)}$, $\boldsymbol{Z}_{(s)}$, $\boldsymbol{Z}_{(o)}$ and $\boldsymbol{Z}_{(d)}$ link each entry in the $\boldsymbol{y}$ vector with the appropriate effect. The levels of the nested trial, location and block effects are sometimes confounded in which case the higher level effect is set to 0. No explicit G×E terms or heterogeneous residual variances were fitted into Eq. (7.3). The expected improvements of these more elaborate models could not be verified because of computational limitations caused by the size of the data set. Furthermore, these models are not handled by the benchmark method described by Bernardo (1994, 1995, 1996a,c) and would therefore exclude an objective comparison.

The covariance matrix $\boldsymbol{G}$ for the random effects in the model can be represented as

$$\boldsymbol{G} = \begin{bmatrix} \boldsymbol{I}\sigma_c^2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{A}_s\sigma_s^2 & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{A}_o\sigma_o^2 & \boldsymbol{0} & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{D}\sigma_d^2 & \boldsymbol{0} \\ \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{0} & \boldsymbol{I}\sigma_r^2 \end{bmatrix} \tag{7.4}$$

The matrices $\boldsymbol{A}_s$ and $\boldsymbol{A}_o$ model the covariance between inbred lines of the ISSS and Iodent heterotic group respectively. Usually the covariance between two hybrids h_{ij} and $h_{i'j'}$, where lines i and i' belong to the ISSS group and lines j and j' belong to the Iodent group, is modelled as (Stuber and Cockerham, 1966)

$$\mathrm{Cov}(h_{ij}, h_{i'j'}) = f_{ii'}\sigma_s^2 + f_{jj'}\sigma_o^2 + f_{ii'}f_{jj'}\sigma_d^2\,, \tag{7.5}$$

where $f_{ii'}$ is the coefficient of coancestry between two inbred lines i and i' of the ISSS heterotic group and $f_{jj'}$ between two inbred lines j and j' of the Iodent group. The coefficient of coancestry can be calculated based on pedigree information (f^{PED}), but also from SSR (f^{BNO}) or AFLP data (f^{JAC}). The three components of Eq. (7.5) allow to construct the matrices $\boldsymbol{A}_s$, $\boldsymbol{A}_o$ and $\boldsymbol{D}$ using the described coefficients of coancestry. These alternative formulations are compared by means of the restricted likelihood of the model given the data, keeping the fixed effects structure constant. The covariance matrix for the checks is assumed to be an identity matrix. If pedigree or marker data were available for these checks, including this information in the covariance matrix $\boldsymbol{G}$ would improve the model fit as proposed by Bernardo (1995). The variance parameters σ_c^2, σ_s^2, σ_o^2, σ_d^2 and σ_r^2 are estimated through REML optimisation by means of the Average Information algorithm as implemented in the software tool ASReml (Gilmour et al., 2002).

The phenotypical value of each hybrid is estimated as the average of its measurements in the data set, albeit with correction for trial, location and block effects. The vector of

corrected phenotypical values is therefore obtained as (Bernardo, 1994, 1995, 1996a,c)

$$\hat{\boldsymbol{y}}_c = (\boldsymbol{Z}'_{(d)}\boldsymbol{Z}_{(d)})^{-1}\boldsymbol{Z}'_{(d)}(\boldsymbol{y} - \mu - \boldsymbol{X}_{(m)}\boldsymbol{m} - \boldsymbol{X}_{(m.l.t)}\boldsymbol{m.l.t} - \boldsymbol{X}_{(m.l.t.b)}\boldsymbol{m.l.t.b})\,. \tag{7.6}$$

The elements of $\hat{\boldsymbol{y}}_c$ are used as a training set for building the ε-SVR prediction model. Apart from training an ε-SVR model we also implemented the prediction system proposed by Bernardo (1994, 1995, 1996a,c) based on Best Linear Prediction (BLP). A validation subset $\hat{\boldsymbol{y}}_{cv}$ of size l' is predicted from the remaining entries $\hat{\boldsymbol{y}}_{ct}$ as

$$\hat{\boldsymbol{y}}_{cv} = \boldsymbol{C}_{vt}\boldsymbol{V}_t^{-1}\hat{\boldsymbol{y}}_{ct}\,. \tag{7.7}$$

$\boldsymbol{C}_{vt}$ is an $l' \times (l - l')$ matrix containing the covariances between validation and training hybrids. $\boldsymbol{V}_t$ is the variance-covariance matrix of the $l - l'$ training hybrids. Elements of $\boldsymbol{C}_{vt}$ and non-diagonal elements of $\boldsymbol{V}_t$ are computed using Eq. (7.5). The i-th diagonal element of $\boldsymbol{V}_t$ is equal to $\sigma_s^2 + \sigma_o^2 + \sigma_d^2 + \frac{\sigma_r^2}{n_i}$, where n_i is the number of records of the i-th hybrid in the training set. The prediction accuracy of BLP is established using a leave-one-out-crossvalidation. This means that each of the 2361 hybrids are individually predicted using a vector $\hat{\boldsymbol{y}}_{ct}$ containing the corrected phenotypical effects of the 2360 remaining hybrids. The algorithm was implemented in C++ using the matrix routines provided in the GNU Scientific Library (Galassi et al., 2009).

7.2.3 ε-insensitive support vector machines regression

As described in Chapter 3, the construction of an ε-SVR prediction model requires a set of training examples, each consisting of a vector of predictors $\boldsymbol{x}$ and an associated response value y. In this case, we want to predict the agronomic performance of a hybrid maize genotype so the elements of the vector $\hat{\boldsymbol{y}}_c$ of Eq. (7.6) are used as response values. For each hybrid, a vector of predictors is constructed based on the molecular fingerprints of its parents. Each element of $\boldsymbol{x}$ represents the expected frequency of a particular allele at a particular locus of that hybrid. For the 75 complete SSR markers, this representation result in a vector $\boldsymbol{x}$ containing 515 allele frequencies. The use of expected instead of actual allele frequencies is necessary because some inbred lines are still heterozygous at one or more loci which prevents the exact deduction of the molecular fingerprints of their children. If for example a hybrid is created from two inbred lines which are both heterozygous at a particular locus, the four elements of its vector $\boldsymbol{x}$ that pertain to the involved alleles are set to 0.25, while the elements representing the other alleles of that particular locus are set to 0. The AFLP markers introduce an additional 569 predictors. As AFLP is a dominant marker type, we are forced to assume complete homozygosity of the inbred lines

to calculate the expected allele frequencies of the marker alleles, which limits the number of possible values for each AFLP-based predictor to 0, 0.5 or 1.

As explained in Chapter 3, the dual formulation of the ε-SVR framework allows to plug in different kernel functions which map the original training examples in a different, generally higher-dimensional space of predictors called the feature space. Not all symmetric functions over $\mathcal{X} \times \mathcal{X}$ are kernels that can be used in a SVM. Since a kernel function K is related to an inner product, it has to satisfy some conditions that arise naturally from the definition of an inner product and are given by Mercer's theorem: the kernel function has to be positive semi-definite (PSD). A commonly used kernel function is the Gaussian kernel defined as

$$K(\boldsymbol{x}, \boldsymbol{z}) = \exp(-\gamma \|\boldsymbol{x} - \boldsymbol{z}\|^2)\,, \tag{7.8}$$

where γ is a kernel specific parameter which allows to find a linear function in an infinitely large feature space (Shawe-Taylor and Cristianini, 2004). Most all-round kernels like the Gaussian or polynomial kernel require the knowledge of one or several additional kernel parameters. The use of context specific kernel functions, however, can avoid the computationally exhausting grid searches needed to identify these parameter values that allow a minimal generalisation error.

Dot products in feature space are in fact measures of similarity between cases, so the use of PSD genetic similarity measures as kernel functions is a valid option. The Jaccard similarity measure of Eq. (7.2) is commonly used when genotyping is based on dominant molecular markers like AFLP. As this similarity measure is PSD (Gower and Legendre, 1986), it can be used as a kernel function in the formulation of an ε-SVR-based prediction model. A useful PSD genetic similarity measure for co-dominant markers is the complement of the Modified Rogers' distance (Rogers, 1972) or MRD defined as

$$s_{ij}^W = 1 - d_{ij}^W \quad \text{with} \quad d_{ij}^W = \frac{1}{\sqrt{2l}} \sqrt{\sum_{k=1}^{l} \sum_{z=1}^{n_k} (p_{kz}^i - p_{kz}^j)^2}\,,$$

where l is the number of genotyped loci, n_k is the number of alleles for locus k and p_{kz}^i, p_{kz}^j represent the allele frequency for the z-th allele of locus k for individual i and j respectively. As demonstrated in Melchinger (1999), there is a linear relationship between ΔH, the panmictic-midparent heterosis and $(d_{kl}^W)^2$ under the assumption of biallelism and absence of epistasis. Therefore, this similarity measure should prove itself useful when used as a kernel function for hybrid prediction.

The weighted sum of two PSD matrices produces a PSD matrix as long as the weights are positive. Computing the weighted sum of different kernel functions therefore creates

a new kernel function. Returning to the concept of a feature space, this operation has the effect of augmenting the dimensions of the feature space related to the first kernel, with the dimensions of the feature space related to the second kernel (Shawe-Taylor and Cristianini, 2004). When we apply this to the Jaccard and MRD kernel functions we have a way to combine SSR and AFLP data into a single regression function. We call the resulting function the Jaccard-MRD kernel.

7.2.4 Cross-validation and grid search

To assess the generalisation error of each ε-SVR model, we rely on a leave-one-out cross-validation procedure. It is, however, infeasible to redo the REML optimisation for each reduced training set as removing records of a randomly chosen hybrid might further unbalance or even disconnect the phenotypic data. In Chapter 5, it was demonstrated that these connectivity issues can seriously bias the estimators and predictors obtained from a linear mixed model analysis. It is therefore assumed that differences between the estimators of the fixed nuisance parameters, calculated using only the data from the training hybrids, and those estimated using all data, are negligible. The reported results therefore do not account for the loss of prediction accuracy in the linear mixed model caused by data reduction.

For each hybrid in the data set, a different ε-SVR model is trained using the corrected phenotypical values ($\hat{\boldsymbol{y}}_c$) of the remaining hybrids as a training set. As explained in Chapter 3, the construction of an ε-regression model from a set of training examples requires values for ε, C and γ when using the Gaussian kernel. Finetuning these variables can greatly improve the generalisation capacity of the prediction system. To find the optimal values for these parameters, a grid search was performed as described by Hsu et al. (2003). During this grid search all combinations of ε, C and if necessary γ were tested for each cross-validation routine, where ε and γ ranged from 2^{-15} to 2^{4} and C ranged from 2^{-5} to 2^{15}. The software libSVM (Chang and Lin, 2006), which allows easy integration of non-standard kernel functions, was used for all regressions. Calculations were performed on a Linux cluster containing 8 nodes, each having 2 Dual-Core Intel R Xeon R CPU 3.00GHz processors, 1Gb of RAM and running a 2.6.5 kernel.

When reporting prediction accuracies, several artificial measures could be used to compare models and techniques. A commonly used measure of prediction accuracy is the root mean squared error (RMSE) defined as the root of the summed squared differences between the actual and the predicted values divided by the number of predictions. Although this measure allows for easy comparison between different models and data sets, it is dependent

on the unit of measurement of the response variable. Comparing accuracies of similar techniques or models on traits measured in a different unit or scale is therefore not possible. Interpreting a RMSE is also quite hard when the reader has no reference for comparing the obtained results. Another commonly used measure is the squared Pearson correlation R^2 between the actual and the predicted value. This coefficient of determination, expressed as a number between 0 and 1, is however dependent on the variance of the predictor variable and resulting predictions. The larger this variance, the larger the obtained correlation will be. This means for example that the correlation between the actual and predicted value for a regression on yield will be larger in natural populations compared to advanced breeding pools with lower yield variance. This property makes it hard to compare published results between prediction methods when different data sets are used. As both criteria seem to cover each others' weaknesses, we compare the different prediction systems by calculating both the R^2 and RMSE values.

7.3 Results

7.3.1 Linear mixed model fit

The average coefficient of coancestry calculated from pedigree data differs substantially from the averages calculated from SSR or AFLP data as can be seen from Table 7.1. Despite the apparent differences between the mean values for f^{PED}, f^{JAC} and f^{BNO}, the Spearman rank correlations between these estimators are moderately high. The AFLP-based coefficients seem to represent an intermediate value between the high SSR- and low pedigree-based coefficients. As can be seen from Table 7.2, the observed correlations between the two marker-based coefficients of coancestry are higher than the correlations between a marker-based and a pedigree-based coefficient for both heterotic groups. However, AFLP-based estimators are closer to the pedigree-based coefficients than the SSR-based alternatives. Apparently all calculated correlations within the Iodent group are greater than those of the ISSS group.

We use the restricted log-likelihood resulting from the REML optimisation process to determine the best fitting covariance structure for Eq. (7.4). Table 7.3 gives an overview of these restricted log-likelihoods for the linear mixed model of Eq. (7.3) where the matrices $\boldsymbol{A}_s$, $\boldsymbol{A}_o$ and $\boldsymbol{D}$ are either considered diagonal or constructed according to Eq. (7.5) using pedigree, SSR or AFLP data for the calculation of the coefficients of coancestry. The models with a non-diagonal covariance matrix for the SCA values always have a lower

Table 7.1: Minimum, maximum and average coancestries based on pedigree (f^{PED}), AFLP (f^{JAC}) and SSR (f^{BNO}) for the two heterotic groups used in this study

		f^{PED}	f^{JAC}	f^{BNO}
Iodent	avg	0.27	0.38	0.45
	min	0	0.04	0.01
	max	0.88	0.99	0.98
ISSS	avg	0.17	0.23	0.31
	min	0	0	0
	max	0.78	0.94	0.95

Table 7.2: Spearman rank correlations between coefficients of coancestry based on pedigree (f^{PED}), AFLP (f^{JAC}) and SSR (f^{BNO}) data for the Iodent and ISSS heterotic groups

ρ	Iodent	ISSS
$f^{\mathrm{PED}} \leftrightarrow f^{\mathrm{JAC}}$	0.79	0.69
$f^{\mathrm{PED}} \leftrightarrow f^{\mathrm{BNO}}$	0.75	0.67
$f^{\mathrm{JAC}} \leftrightarrow f^{\mathrm{BNO}}$	0.90	0.77

Table 7.3: Log likelihoods for the linear mixed model of Eq. (7.3) with fixed nuisance factors but different formulations for $\boldsymbol{G}$. The covariance matrices for GCA and SCA effects are either diagonal, based on pedigree, SSR or AFLP data.

$\boldsymbol{A}$	$\boldsymbol{D}$	yield	moisture %	flowering
diagonal	diagonal	−588609	−201915	−158515
pedigree	diagonal	−588590	−201872	−158498
pedigree	pedigree	−588659	−202056	−158571
SSR	diagonal	−588585	−201879	−158504
SSR	SSR	−588681	−202142	−158600
AFLP	diagonal	**−588583**	**−201855**	**−158487**
AFLP	AFLP	−588639	−201961	−158537

likelihood than their diagonal counterparts. This means that the covariance between SCA values should be modelled as 0 as it seems to fit better than the product of both coefficients of coancestry as in Eq. (7.5).
The model with AFLP-based $\boldsymbol{A}$ matrices and an identity $\boldsymbol{D}$ matrix results in the highest restricted log-likelihood for all traits under study. A model with SSR-based $\boldsymbol{A}$ matrices and a diagonal $\boldsymbol{D}$ matrix gives the second highest restricted log-likelihood for yield, but performs worse than the pedigree-based equivalent for moisture content and days until flowering. These results indicate that the AFLP-based coefficient of coancestry approximates better the actual relatedness between hybrids compared to the pedigree-based and even the SSR-based coefficient for this data set. All subsequent regressions and predictions are therefore based on the results of the linear mixed models with AFLP-based $\boldsymbol{A}$ and diagonal $\boldsymbol{D}$ matrices in Eq. (7.4).

7.3.2 ε-SVR

When testing new hybrid prediction algorithms, the main interest lies in the estimation of the total genetic value of untested hybrids. We use the corrected phenotypical values in vector $\hat{\boldsymbol{y}}_c$ from Eq. (7.6) as a training set for building a regression model. By means of the standard leave-one-out cross-validation strategy, the predictive capabilities of the different

kernels are compared to each other. Table 7.4 gives an overview of the obtained prediction accuracy for the different combinations of trait, marker type and kernel functions. The last column represents the leave-one-out cross-validation accuracy of the prediction by means of Eq. (7.7) using marker based coefficients of coancestry to model $\boldsymbol{C}^{vt}$ and $\boldsymbol{V}^{t}$.
When the molecular information is restricted to microsatellite data, the ε-SVR-based models, albeit with a minimal difference, provide better prediction accuracies than BLP. Comparing the three kernel functions, we notice that the two non-linear kernel functions always perform slightly better than the linear one. This observation demonstrates the advantage of performing a linear regression in a kernel induced feature space. The similarity based MRD kernel function performs just as good as the Gaussian kernel but does not require the finetuning of an additional kernel parameter so using MRD to build an optimised prediction model takes far less computation time. Prediction accuracies of ε-SVR and BLP are also very similar when the molecular fingerprints of the inbred lines are restricted to AFLP markers. For yield and days until flowering, ε-SVR is slightly superior, while BLP is preferred for moisture content. Again the non-linear kernels perform better than their linear counterpart and the parameter free Jaccard-based kernel function provides a valid alternative to the Gaussian kernel.
When we need to decide between SSR- and AFLP-based features, we notice that for each trait under study, the AFLP markers provide equal or slightly better prediction accuracies than the SSR markers. Examining the restricted log-likelihood of the linear mixed model revealed the same preference for the dominant AFLP marker data. In either case the differences are minimal to say the least so these conclusions should not be generalised to other data sets. For all three traits, combining the information of SSR and AFLP markers provides the highest prediction accuracy over all applied methods but the gain in precision is minimal as both sets of markers seem to be equally informative in this case.
The maximum obtained squared Pearson correlations using an ε-SVR based model are 0.34, 0.72 and 0.40 for yield, moisture content and days until flowering respectively while these are 0.33, 0.71 and 0.38 for BLP. We can therefore conclude that ε-SVR predictions are at least as accurate as the corresponding analyses using BLP. It should, however, be clear that the reported accuracies for both prediction frameworks need much improvement to allow for a reliable genomic selection, especially for the traits grain yield and days until flowering. The flexibility of the ε-SVR framework provides several opportunities that might enable such improvements, for example by means of feature selection methods. The gradient descent based R2W2 technique described by Weston et al. (2000) and the greedy Recursive Feature Elimination or RFE described by Guyon et al. (2002) are examples of such methods that allow for the identification of markers that have little or no contribution

Table 7.4: Standard leave-one-out prediction accuracies, expressed as squared Pearson correlations and RMSEs (between brackets), on corrected phenotypical values for yield, moisture content and days until flowering. The results are presented according to the type of features (SSR, AFLP or both) and the type of kernel function used during the analysis. The last column represents the accuracy of the predictions obtained with BLP (Bernardo, 1994, 1995, 1996a,c). The prediction method with the highest correlation and lowest RMSE is typesetted in bold for each trait.

SSR				
	linear	Gaussian	MRD	BLP
yield	0.31 (6.80)	0.33 (6.67)	0.33 (6.66)	0.33 (6.72)
moist. cont.	0.69 (1.19)	0.70 (1.16)	0.71 (1.14)	0.71 (1.16)
flowering	0.38 (1.18)	0.39 (1.16)	0.40 (1.16)	0.38 (1.18)
AFLP				
	linear	Gaussian	Jaccard	BLP
yield	0.31 (6.79)	0.34 (6.66)	0.32 (6.76)	0.33 (6.72)
moist. cont.	0.69 (1.18)	0.71 (1.15)	0.70 (1.16)	0.71 (1.13)
flowering	0.38 (1.18)	0.40 (1.16)	0.38 (1.18)	0.38 (1.17)
AFLP+SSR				
	linear	Gaussian	Jaccard-MRD	BLP
yield	0.31 (6.8)	**0.34** (**6.63**)	0.33 (6.68)	-
moist. cont.	0.69 (1.18)	0.71 (1.14)	**0.72** (**1.13**)	-
flowering	0.38 (1.18)	**0.40** (**1.15**)	0.40 (1.16)	-

to the prediction model. Besides the advantage of identifying key markers which could be used as a starting point for more detailed association studies, it is to be expected that removing the useless features shall improve the obtained prediction accuracies; however, further study is required to ascertain this point. Another possible road to improvement is to design specific kernel functions for hybrid prediction. This allows to encode prior knowledge of the learning task into the feature space in which the regression takes place. The advantages of engineering a case-specific kernel function are exemplified by Zien et al. (2000) who designed a kernel for the identification of translation initiation sites in DNA code which resulted in a significantly improved recognition performance compared to the standard kernel functions.

7.4 Discussion

BLP is currently one of the best known methods for the prediction of the phenotypical performance of maize hybrids originating from crosses between unrelated lines, as is the case for most of today's commercial hybrids. We evaluated the use of ε-insensitive Support Vector Machine Regression, as an alternative to BLP, on a real maize breeding data set from the private breeding company RAGT R2n. The idea is to train the ε-SVR algorithm to directly predict the phenotypical values of maize hybrids based on the molecular marker scores of both parental inbred lines and compare the obtained prediction accuracies with those of BLP. The field trial data resulting from a commercial breeding programme are typically very unbalanced and therefore linear mixed modelling is used to adjust the phenotypical measures for location, trial and block effects. For each hybrid, the average of the corrected plot measurements for yield, grain moisture contents and days until flowering are used as predictands while the AFLP- and SSR-based molecular fingerprints of the parental inbred lines serve as predictor variables.

We calculated the coefficients of coancestry based on pedigree, SSR and AFLP data for all pairwise combinations of inbred lines within each of the two heterotic groups. The Spearman rank correlations between the obtained similarity measures are moderately high but the marker-based coefficients generally indicate a higher level of relatedness between the individual lines. This discrepancy might be explained by the unequal parental contributions that can occur after several generations of inbreeding during line development. A standard pedigree analysis is not able to detect these shifts and assumes equal contributions from both parents. Another possible cause of bias is the assumption of unrelated ancestor individuals which is often impossible to verify. As the described deviations are higher for

the ISSS lines, we can assume that these departures from theoretical assumptions are more pronounced within this heterotic group.

The resulting likelihood of the REML procedure for estimating the variance components of the linear mixed model allows to identify the best fitting covariance structure. For the data set at hand, the likelihood of the model with a diagonal covariance matrix $\boldsymbol{D}$ for the SCA effects is higher than the pedigree-, AFLP- and SSR-based alternatives. This result has also been observed in other data sets (Piepho H.P., 2006 personal communication at the session "BLUP in Plant Breeding", XIII EUCARPIA Biometrics in Plant Breeding Section Meeting, Zagreb, Croatia) and demonstrates that the base assumptions underlying the derivation of Eq. (7.5) in Stuber and Cockerham (1966), in particular the absence of linkage disequilibrium and different effects of the same alleles in the two populations, do not hold in an advanced breeding pool. Moreover Eq. (7.5) is a simplification, leaving out all interaction terms besides the dominance effect and therefore assuming that epistasis is negligible. We also noticed that using products of coefficients of coancestry as entries in $\boldsymbol{D}$ does not guarantee a positive definite covariance matrix for the SCA values which is counterintuitive and can lead to convergence problems of the REML algorithm.

The AFLP-based coefficient of coancestry is preferred when modelling the covariance between the GCA effects of the parental inbred lines although the restricted log-likelihood of a model using SSR-based coancestries is comparable. This observation seems to contradict the results obtained in Chapter 6, where the pedigree-based coancestry estimator resulted in the highest restricted log-likelihood for the traits grain moisture content and days until flowering. Both studies, however, differ substantially with respect to the fitted linear mixed model which might be sufficient to explain this observed discrepancy. Marker similarities are corrected for the difference between identity in state and identity by descent by means of the average marker similarity of each inbred line with all inbred lines of the complementary heterotic group. As indicated by Bernardo (1996c), this approach assumes homogeneous allele frequencies among these heterotic groups. As this was generally not the case for the Iodent and ISSS group in this study, the presented coefficients of coancestry are biased. It is to be expected that the model fit will improve when the marker-based coefficients of coancestry are derived from estimators of parental contribution as described in Bernardo et al. (2000). Unfortunately, the elaborate pedigree of the 197 selected inbred lines does not allow the fingerprinting of all ancestral individuals to calculate these parental contributions from SSR or AFLP similarities. This will generally be the case when working with historically evolved heterotic groups.

By using the most likely linear mixed model, we can correct the phenotypical values for each hybrid for nuisance factors and use these estimators as a training set for the construction of

an ε-SVR model. Correlations between real and predicted phenotypical values, obtained by means of a leave-one-out cross-validation, show that the non-linear kernels perform better than their linear counterpart for every combination of trait and marker type. This demonstrates the advantage of performing a linear regression in a kernel induced feature space. These non-linear kernels generally allow to match or slightly improve the accuracy of the currently best performing prediction method for crosses between unrelated inbred lines. The training of an ε-SVR model does, however, assume the knowledge of several parameters like the width ε of the insensitivity tube, the error weighting variable C and possibly one or several kernel function parameters. These parameters can be optimised by a simple grid search in combination with cross-validation routines but this can become computationally exhausting when the number of required kernel parameters is large. Subject-specific kernel functions like the presented Jaccard measure, MRD and their linear combinations can avoid the necessity of extra kernel parameters while allowing similar prediction accuracies. Both the Jaccard measure and the complement of the modified Rogers' distance are PSD similarity measures and therefore represent a dot product in some feature space. In practice, the requirement of a kernel function to be PSD turns out to be a very strict assumption. Several references can be found where a symmetric non-PSD similarity function is used within the standard SVM framework as a heuristic approach (Bahlmann et al., 2002; Decoste and Schölkopf, 2002; Haasdonk and Keysers, 2002). Problems like nonconvexity of the optimisation problem can be handled by adding an additional term to the objective function of Eq. (3.17) as described in Fan et al. (2005). This approach guarantees that the optimisation process converges to a stationary point but only in the case of a PSD kernel function this point is the unique optimal value. This information leads one to suspect that several other similarity measures, PSD or not, and their linear combinations could increase the prediction accuracy of ε-SVR models but further study is obviously needed to ascertain this. Another advantage of the ε-SVR methodology is the easy integration of different types of molecular and even descriptive morphological data as features. As there is no straightforward way to incorporate all this information into the covariance matrices of BLP, ε-SVR allows for a greater flexibility when the prediction system has to be implemented into an existing breeding programme. Easy feature selection heuristics like the greedy Recursive Feature Elimination (Guyon et al., 2002) should allow for the identification of specific molecular markers and possibly parental morphological properties that are crucial for the construction of the prediction model. When evaluating new inbred lines one can make the trade-off between the cost of collecting a certain feature and the increase in prediction accuracy that this feature represents.

To conclude we can state that, although further comparisons using other data sets are

necessary, the presented ε-SVR models can generally compete with BLP. Parameter optimisation, feature selection algorithms and problem-specific kernel functions are several promising aspects of this recent technique which need further investigation in the context of hybrid prediction.

CHAPTER 8

Support Vector Machine regression versus Best Linear Prediction

8.1 Introduction

The prediction of phenotypic performance from molecular marker data receives increasing attention from plant breeders, as the cost of phenotyping is gradually overtaking the cost of genotyping (Bernardo, 2008). In this field of research, plant species for which it is relatively easy to create and cross almost fully homozygous inbred lines, are particularly useful as they allow to study the effect of a single gamete in different genetic backgrounds. In Chapter 7, data from a commercial maize breeding programme was used to compare the phenotypic prediction accuracy of ε-insensitive Support Vector Machine Regression (ε-SVR) to that of the method advocated by Bernardo (1994, 1995, 1996a,c) based on Best Linear Prediction (BLP). The reported prediction accuracies, determined by means of a leave-one-out cross-validation routine, indicate that both methods are equally good at predicting phenotypes for three important agronomic traits. In this chapter, we further examine several key aspects of hybrid prediction by means of ε-SVR and BLP which allows to clarify the strengths and weaknesses of both methods.

Field trial data originating from commercial hybrid breeding programmes is typically very unbalanced. Tester lines are parents of many hybrids, while other inbred lines may appear only once in the company's pedigree. Furthermore, there is usually quite a substantial

This chapter has been redrafted after Maenhout S., Haesaert G. and De Baets B. (2010). Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction *Theoretical and Applied Genetics*, 120:415-427.

difference in the number of field trials in which a promising hybrid is tested compared to the often single trial results of the lesser candidates. Both ε-SVR and BLP require a set of hybrids for which a molecular fingerprint and a single response value for each trait are available. Such a phenotypic response value or score can be obtained by means of a linear mixed model analysis of the unbalanced phenotypic data, but different model assumptions and prediction approaches can lead to very different results. We study the impact of these assumptions by comparing three different data preparation methods. In the linear mixed models described by Bernardo (1994, 1995, 1996a,c) and used in Chapter 7, the non-genetic effects of growing seasons, locations and blocks are assumed to be fixed while the genotypic and G×E effects are assumed to be random. Bernardo (1994, 1995, 1996a,c) obtains a single phenotypic score for a particular hybrid by taking the average of all its phenotypic measurements, after correcting them by means of the estimated fixed effects. One could, however, just aggregate the BLUPs of the genotypic components directly to obtain a single score for each hybrid. Besides these two data preparation methods, we also study a third approach in which the genotypic effects are assumed to be fixed while the non-genetic nuisance parameters are treated as random.

In Chapter 7, we used all hybrids that are represented in the available unbalanced phenotypic data and the entire set of genotyped molecular markers to compare the prediction accuracy of ε-SVR and BLP. The sensitivity of both methods to a reduction in the number of training examples or genotyped molecular markers is however left unexamined. To assess the impact of the training sample size and marker information content on the prediction accuracy, we apply both methods to selected subsets of the training sample and molecular marker fingerprint. The results allow to identify minimum sample size requirements of ε-SVR and BLP models that are trained using comparable, unbalanced data sets.

The accuracy of hybrid prediction techniques is generally measured by some form of cross-validation strategy (Bernardo, 1994, 1995, 1996a,c; Charcosset et al., 1998; Schrag et al., 2007, 2009). Schrag et al. (2007) argue that an assessment of prediction accuracy by means of a leave-one-out cross-validation routine does not reflect practical breeding circumstances where a new inbred line would be crossed with only a few tester lines from the opposite heterotic group. They propose a modified cross-validation sampling scheme that requires a mating design in which every inbred line from one heterotic group is crossed with all lines belonging to the complementary heterotic group. To allow for such a realistic assessment of prediction accuracy in an unbalanced setting, Bernardo (1996c) uses cross-validation schemes that simulate a lack of prior information on one or both parental inbred lines of a newly created hybrid. Although these schemes represent an improvement, they do not solve the fundamental problem of cross-validation-based accuracy measures. As the train-

ing examples are predicted marginal to the effects of growing seasons, test locations and possibly fertiliser or irrigation treatments, the resulting cross-validation-based prediction accuracy measures do not take into account the extra level of uncertainty that is caused by G×E effects (Welham et al., 2004). This implies that the observed correlation between the predicted marginal genotypic values and those estimated conditional on a specific level of the environmental factors (i.e. in an additional field trial in a specific year and geographical region) might differ substantially from the cross-validation-based prediction accuracy. To quantify this expected discrepancy, we performed a validation field trial using 49 hybrids which were created by crossing 7 Iodent lines with 7 Iowa Stiff Stalk Synthetic (ISSS) lines. The phenotypic performance of these hybrids was measured in a multi-environment trial at three locations in the South of France. Prediction accuracy is determined by correlating the resulting estimates for total genotypic value and SCA to the predictions of ε-SVR and BLP models, constructed from the unbalanced training data.

To summarise, we recapitulate the three main objectives of the research presented in this chapter: (1) to identify the best method for distilling a single phenotypic score for each hybrid in an unbalanced data set, (2) to compare the prediction accuracy of ε-SVR and BLP when the sample size and information content of the molecular marker fingerprint are reduced and (3) to compare the prediction accuracy measures obtained through various cross-validation schemes with those obtained by means of a validation field trial.

8.2 Material and methods

To achieve the three objectives, this study investigates the impact of changing the levels of the factors influencing them which are summarised in Table 8.1 and discussed below.

8.2.1 Training data analysis

The data used in this study is a subset of the genotypic and phenotypic information generated by the grain maize breeding programme of the private company RAGT R2n, and is described in detail in Chapter 4. We use a slightly different data subset selection procedure resulting in 40432 phenotypic measurements on 2354 hybrids originating from unbalanced crosses between 92 Iodent and 105 ISSS lines. We study the traits grain yield, grain moisture content and days until flowering, which were measured in 1280 multi-environment trials representing 110 locations spread over Europe from 1989 to 2005. The 197 parental inbred lines are genotyped with 101 SSR markers, which are evenly distributed

Table 8.1: Overview of the different traits, training data preparation methods, molecular marker-based predictors, prediction methods, sampling schemes and methods for prediction accuracy measurement that are combined in this chapter.

factor	levels
trait	grain yield grain moisture content days until flowering
training data preparation	($\hat{\boldsymbol{y}}_T^{rp}$) random phenotypes ($\hat{\boldsymbol{y}}_T^{rg}$) random genotypes ($\hat{\boldsymbol{y}}_T^{fg}$) fixed genotypes ($\hat{\boldsymbol{d}}_T^{rs}$) random SCA
predictor	AFLP SSR
prediction method	ε-SVR BLP
sampling scheme	random sampling test-cross sampling new-cross sampling random marker reduction
prediction accuracy measurement	cross-validation validation field trial: ($\hat{\boldsymbol{y}}_V^{rg}$) random genotypes ($\hat{\boldsymbol{d}}_V^{rs}$) random SCA

over the maize genome according to the proprietary linkage map of RAGT R2n. Due to problems identifying some SSR alleles (null alleles), only 75 markers, which have a complete profile over all inbred lines, are used. AFLP fingerprints are generated using 11 *Pst*I-*Mse*I and 4 *Eco*RI-*Mse*I primer combinations producing 569 polymorphic bands in total.
The construction of an ε-SVR or BLP prediction model for a specific quantitative trait requires a single response value for each training example representing the genetic potential of each genotype at each location and year. We consider three methods of constructing such a response value based on linear mixed modelling of the trial data. We also predict SCA values from a mixed model analysis.

Random phenotypes

In the first approach, we consider the environmental effects (e.g. year, location, block, ...) as fixed effects, while we consider GCA, SCA and all G×E interactions as random effects. A detailed description of this linear mixed model for the three traits under study can be found in Chapter 6. The variance structures of GCA and SCA effects are modelled according to Stuber and Cockerham (1966) where we use the AFLP fingerprints to obtain estimators for the pairwise coefficient of coancestry between inbred lines i and j belonging to the same heterotic group as (Bernardo, 1993)

$$f_{ij} = \frac{f_{ij}^{\mathrm{JAC}} - \frac{1}{2}(\bar{f}_{i.}^{\mathrm{JAC}} + \bar{f}_{j.}^{\mathrm{JAC}})}{1 - \frac{1}{2}(\bar{f}_{i.}^{\mathrm{JAC}} + \bar{f}_{j.}^{\mathrm{JAC}})}\,, \tag{8.1}$$

where f_{ij}^{JAC} is the Jaccard similarity coefficient between the AFLP fingerprints of lines i and j. $\bar{f}_{i.}^{\mathrm{JAC}}$ is the average Jaccard similarity coefficient between inbred line i and all lines belonging to the opposite heterotic group. As shown in Chapter 7, this estimator for the coefficient of coancestry resulted in the highest restricted log-likelihood, when compared to several other estimators that use pedigree, AFLP or SSR marker information. The genotypic estimate is obtained by averaging over all measurements of a single hybrid in the response vector $\boldsymbol{y}$ after correction for the estimated fixed environmental effects as

$$\hat{\boldsymbol{y}}_T^{rp} = (\boldsymbol{Z}'\boldsymbol{Z})^{-1}\boldsymbol{Z}'(\boldsymbol{y} - \boldsymbol{X}\hat{\boldsymbol{\beta}})\,,$$

where $\boldsymbol{Z}$ is a design matrix linking the phenotypic measurements in vector $\boldsymbol{y}$ to each hybrid in vector $\hat{\boldsymbol{y}}_T^{rp}$. Vector $\hat{\boldsymbol{\beta}}$ contains the estimated effects for the levels of each nuisance factor and these are linked to the response vector $\boldsymbol{y}$ by means of the design matrix $\boldsymbol{X}$. Bernardo (1994, 1995, 1996a,c) calls the entries in vector $\hat{\boldsymbol{y}}_T^{rp}$ phenotypes, as these are not corrected for G×E interaction effects or residual error. The superscript rp is shorthand for random

phenotypes, while the subscript T indicates that this vector was obtained from the training data.

Random genotypes

The second method is to sum the appropriate GCA and SCA BLUPs obtained from the afore mentioned linear mixed model analysis as

$$\hat{\boldsymbol{y}}_T^{rg} = \boldsymbol{Z}_s\hat{\boldsymbol{a}}_s + \boldsymbol{Z}_o\hat{\boldsymbol{a}}_o + \hat{\boldsymbol{d}}_T^{rs}\,, \tag{8.2}$$

where $\hat{\boldsymbol{a}}_s$ and $\hat{\boldsymbol{a}}_o$ are vectors containing BLUPs of the GCA values of lines belonging to the ISSS and Iodent heterotic groups, respectively. The design matrices $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_o$ link each hybrid to the appropriate parental inbred lines. Vector $\hat{\boldsymbol{d}}_T^{rs}$ contains a BLUP of the SCA value for each hybrid. As we treat the GCA and SCA effects as random model factors, we use the superscript rg to indicate this random nature of the genotypic values in vector $\hat{\boldsymbol{y}}_T^{rg}$. This approach implicitly produces genotypic scores that are marginal to all environmental factors in the model such as growing seasons and locations. These marginal scores have larger standard errors compared to estimators that are conditional on one or more environmental factors (Welham et al., 2004), but we prefer them here as they do not require knowledge of the future environmental conditions in which the predicted hybrids will be grown.

Fixed genotypes

The third method of forming a genotypic response is from a linear mixed model with genotypes fixed and non-genetic effects fitted as random. This approach allows to obtain a vector of estimated genotypic fixed effects $\hat{\boldsymbol{y}}_T^{fg}$ without making prior assumptions on the covariance structure of the GCA and SCA components.

Random SCA

Besides training on genotypic or phenotypic scores, we also construct prediction models for the SCA values in vector $\hat{\boldsymbol{d}}_T^{rs}$ of Eq. (8.2).

8.2.2 Validation data

Data description

7 ISSS and 7 Iodent lines were selected from the initial set of 197 inbred lines and pairwise intermated to produce 49 cross-heterotic hybrids. For these hybrids and an additional

6 check varieties, the traits grain yield, grain moisture content and days until flowering were measured in a balanced field trial at three locations in the South of France during the growing season of 2008. The initial selection of 14 parental inbred lines was based on the ε-SVR and BLP predictions of all 9660 possible hybrids between the 105 ISSS and 92 Iodent lines. A greedy search heuristic was used to approach the optimal selection of 14 parental inbred lines such that the ε-SVR and BLP predictions of the 49 hybrids show the largest variance in grain yield. However, several lines in this initial selection were replaced by other lines so that all hybrids had a comparable maturity index. Only 11 of the 49 hybrids were in fact new combinations, while the other 38 already had phenotypic records in the training data. Regardless of potential seed availability, each of the 49 crosses were (re)created under the exact same circumstances, as to avoid non-genetic seed quality differences. At each location of the trial, the 55 hybrids are laid out as a two-replicate resolvable row-column design with 22 rows and 5 columns.

Data analysis

A linear mixed model analysis is performed assuming location effects as fixed and all genetic components and G×E interactions as random. The description of the statistical model follows the notation of Smith et al. (2001) where the vector of phenotypic measurements $\boldsymbol{y}$ is decomposed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\tau} + \boldsymbol{Z}_g\boldsymbol{g} + \boldsymbol{Z}_u\boldsymbol{u} + \boldsymbol{e}\,, \tag{8.3}$$

and $\boldsymbol{\tau}$ is a vector of fixed effects containing main location effects and location-specific effects correcting for extraneous field variation. $\boldsymbol{g} = (\boldsymbol{g}_1', \boldsymbol{g}_2', \boldsymbol{g}_3')'$ is a vector containing the random effects of the 55 hybrids in each of the 3 locations with an associated design matrix $\boldsymbol{Z}_g$. $\boldsymbol{u}$ is also a vector of random effects modelling for location-specific blocking factors. The vector of residuals $\boldsymbol{e} = (\boldsymbol{e}_1', \boldsymbol{e}_2', \boldsymbol{e}_3')'$ is partitioned in three subvectors corresponding to the three locations. For the trait grain moisture contents, the values in vector $\boldsymbol{y}$ were logit transformed.
The vector of genetic effects $\boldsymbol{g}$ is partitioned as

$$\boldsymbol{Z}_g\boldsymbol{g} = \boldsymbol{Z}_c\boldsymbol{c} + \boldsymbol{Z}_s\boldsymbol{a}_s + \boldsymbol{Z}_o\boldsymbol{a}_o + \boldsymbol{Z}_d\boldsymbol{d}\,, \tag{8.4}$$

where $\boldsymbol{c} = (\boldsymbol{c}_1', \boldsymbol{c}_2', \boldsymbol{c}_3')'$ represents a vector containing the genetic effects of the 6 check varieties at each of the three locations, vectors $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$ contain the GCA effects of the parental inbred lines belonging to the ISSS and Iodent heterotic groups respectively and vector $\boldsymbol{d}$ contains the SCA effects of the 49 hybrids at each location. The design matrices

$\boldsymbol{Z}_c$, $\boldsymbol{Z}_s$, $\boldsymbol{Z}_o$ and $\boldsymbol{Z}_d$ separate check and non-check entries and matrices $\boldsymbol{Z}_s$ and $\boldsymbol{Z}_o$ have the additional function of linking the appropriate parental inbred lines to each non-check hybrid in vector $\boldsymbol{g}$. The four random vectors $\boldsymbol{c}$, $\boldsymbol{a}_s$, $\boldsymbol{a}_o$ and $\boldsymbol{d}$ are assumed to be mutually independent. Furthermore, for each of these vectors $\boldsymbol{h} \in \{\boldsymbol{c}, \boldsymbol{a}_s, \boldsymbol{a}_o, \boldsymbol{d}\}$ we assume that the variance has the separable form

$$\mathrm{Var}(\boldsymbol{h}) = \boldsymbol{G}_e \otimes \boldsymbol{G}_v \,, \tag{8.5}$$

where $\otimes$ denotes the Kronecker product. $\boldsymbol{G}_e$ represents a 3×3 symmetric matrix containing the covariance between environments while $\boldsymbol{G}_v$ represents the covariance between the specified genetic components of the validation trial entries. We start by fitting a completely unstructured variance matrix for $\boldsymbol{G}_e$ while assuming an identity matrix for $\boldsymbol{G}_v$. In subsequent steps, the number of REML estimated variance components is reduced by fitting more parsimonious variance models for $\boldsymbol{G}_e$ using restricted maximum likelihood ratio tests in case of comparisons between nested models, or Akaike's Information Criterion (AIC) otherwise. We attempt to fit a first-order factor analytic variance model such that $\boldsymbol{G}_e = \boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}$ where $\boldsymbol{\lambda}$ is a vector of factor loadings and the matrix $\boldsymbol{\Psi}$ is a diagonal matrix containing three location-specific variances (Smith et al., 2001). To obtain a more parsimonious model, the specific variances were sometimes made equal or zero (giving perfect correlation), and/or the loadings made equal (giving a common covariance (Cullis et al., 1998)). In a subsequent reduction, the variances on the diagonal are set equal which results in a compound symmetry model. The simplest model for $\boldsymbol{G}_e$ assumed zero covariance and equal variances.

Once the most parsimonious model for $\boldsymbol{G}_e$ is determined, we try different formulations for $\boldsymbol{G}_v$. We fit an identity matrix for the variance model of the six check varieties in vector $\boldsymbol{c}$ as no molecular marker or pedigree information is available for these varieties. For the vectors $\boldsymbol{a}_s$ and $\boldsymbol{a}_o$, containing the GCA effects of the inbred lines, we try to fit the different coefficient of coancestry derived matrices $\boldsymbol{A}$ described in Chapter 6 or an identity matrix. In a similar way, we compare the different coefficient of fraternity-based matrices $\boldsymbol{D}$ for the variance matrix $\boldsymbol{G}_v$ pertaining to the vector $\boldsymbol{d}$. Sometimes, the most parsimonious model is obtained by not using the separable form of Eq. (8.5) but directly fitting a common GCA or SCA effect for all three locations.

The variance of each vector of residuals $\boldsymbol{e}_i$ that make up vector $\boldsymbol{e}$ in Eq. (8.3) is modelled as a separable process in the direction of rows and columns so we can write $\mathrm{Var}(\boldsymbol{e}_i) = \boldsymbol{\Sigma}_{ic} \otimes \boldsymbol{\Sigma}_{ir}$ where $\otimes$ denotes the Kronecker product. The matrices $\boldsymbol{\Sigma}_{ic}$ and $\boldsymbol{\Sigma}_{ir}$ are either identity matrices or contain first order autoregressive correlations to account for spatial variation as described in A. Gilmour and Verbyla (1997), Smith et al. (2001) and Oakey et al. (2007).

Table 8.2 gives an overview of the final model for the variance structure of vectors $\boldsymbol{g}$ and $\boldsymbol{e}$ for each trait.

Table 8.2: Summary of the variance structures fitted on the measurements of the validation data set for the traits grain yield, grain moisture content and days until flowering.

component	yield	moisture content	flowering
Var($\boldsymbol{c}$)	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \boldsymbol{I}_6$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \boldsymbol{I}_6$	CS
Var($\boldsymbol{a}_s$)	CS	CS	$(\boldsymbol{\Lambda}\boldsymbol{\Lambda}' + \boldsymbol{\Psi}) \otimes \boldsymbol{I}_7$
Var($\boldsymbol{a}_o$)	$\boldsymbol{I}_3 \otimes \boldsymbol{A}_o$	$\boldsymbol{A}_o$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}') \otimes \boldsymbol{I}_7$
Var($\boldsymbol{d}$)	$\boldsymbol{D}$	$(\boldsymbol{\lambda}\boldsymbol{\lambda}' + \boldsymbol{\Psi}) \otimes \boldsymbol{D}$	$\boldsymbol{D}$
Var($\boldsymbol{e}_1$)	$\boldsymbol{I}_5 \otimes$ AR1	$\boldsymbol{I}_5 \otimes \boldsymbol{I}_{11}$	$\boldsymbol{I}_5 \otimes \boldsymbol{I}_{11}$
Var($\boldsymbol{e}_2$)	$\boldsymbol{I}_5 \otimes$ AR1	AR1 $\otimes \boldsymbol{I}_{11}$	AR1 $\otimes \boldsymbol{I}_{11}$
Var($\boldsymbol{e}_3$)	$\boldsymbol{I}_5 \otimes$ AR1	$\boldsymbol{I}_5 \otimes$ AR1	AR1 $\otimes \boldsymbol{I}_{11}$

$\boldsymbol{\lambda}$: loadings of a first order factor analytic covariance model
$\boldsymbol{I}_l$: identity matrix of size l
CS: compound symmetry, common genetic covariance over locations
$\boldsymbol{\Psi}$: site specific variances of a factor analytic covariance model
AR1: first order autoregressive covariance
$\boldsymbol{A}_o$: matrix containing coancestry coefficients of Iodent lines according to Eq. (8.1). In case $\boldsymbol{A}_o$ is not part of a Kronecker product, a common Iodent GCA effect for all three locations was fitted
$\boldsymbol{D}$: matrix containing fraternity coefficients for the 49 hybrids according to Stuber and Cockerham (1966). In case $\boldsymbol{D}$ is not part of a Kronecker product, a common SCA effect for all three locations was fitted

The linear mixed model analysis of the validation trial data provides BLUPs for the GCA and SCA components which are summed according to Eq. (8.2) to obtain an estimate of the genotypic value for each of the 49 hybrids. These estimates are grouped in the vector $\hat{\boldsymbol{y}}_V^{rg}$ where the subscript V indicates their validation trial origin. The vector $\hat{\boldsymbol{d}}_V^{rs}$ contains the BLUPs of the 49 SCA values.

8.2.3 Prediction methods

ε-Insensitive Support Vector Machines Regression

In Chapter 7, it was shown how ε-SVR can be used to predict the phenotypic performance of new hybrids using unbalanced phenotypic training data and AFLP or SSR marker fingerprints as predictors. Cross-validation results indicated that solving the linear regression problem in an infinite-dimensional space by means of a Gaussian kernel function results in a higher prediction accuracy compared to a linear solution in the original input space. Using the Gaussian kernel function requires a value for the kernel parameter γ and the optimisation function that is minimised during the construction of an ε-SVR prediction model requires two additional parameters C and ε. In this study, optimal values for C, ε and γ are found by an expensive grid-search over this three-dimensional space with a v-fold cross-validation prediction accuracy as optimisation criterion. To reduce the computational effort, the ε-SVR parameter searches in the present study were guided by the Efficient Global Optimisation or EGO algorithm reported by Jones et al. (1998). The criterion to be optimised was the squared Pearson correlation coefficient obtained by a v-fold cross-validation where $v = 20$.

Best Linear Prediction

Bernardo (1994, 1995, 1996a,c) makes predictions for a set of single crosses as

$$\hat{\boldsymbol{y}}_P = \boldsymbol{C}_{PT}\boldsymbol{V}_T^{-1}\hat{\boldsymbol{y}}_T\,, \tag{8.6}$$

where $\boldsymbol{C}_{PT}$ is the genetic covariance matrix between the hybrids in the training set $\hat{\boldsymbol{y}}_T$ and the hybrids to be predicted and $\boldsymbol{V}_T = \mathrm{Var}(\hat{\boldsymbol{y}}_T)$ is the variance matrix of the hybrids in the training set. The genetic covariances in the matrices $\boldsymbol{C}_{PT}$ and $\boldsymbol{V}_T$ are obtained from a simplification of the covariance model described in Stuber and Cockerham (1966)

$$\mathrm{Cov}(h_{ij}, h_{i'j'}) = \theta_{ii'}\sigma_s^2 + \theta_{jj'}\sigma_o^2 + \theta_{ii'}\theta_{jj'}\sigma_d^2\,,$$

where h_{ij} and $h_{i'j'}$ are two hybrids for which the parental inbred lines i and i' belong to the ISSS heterotic group and the lines j and j' belong to the Iodent group. $\theta_{ii'}$ and $\theta_{jj'}$ are the coefficients of coancestry estimated from SSR (Bernardo, 1993) or AFLP marker information, the latter based on Eq. (8.1). The additive variance parameters σ_s^2 and σ_o^2 and the dominance variance σ_d^2 are obtained from the REML analysis of the training data. We obtain $\hat{\boldsymbol{y}}_P$ from Eq. (8.6) by solving the system of linear equations

$$\boldsymbol{V}_T\boldsymbol{x}_T = \hat{\boldsymbol{y}}_T \tag{8.7}$$

for $\boldsymbol{x}_T$ through a Cholesky decomposition of $\boldsymbol{V}_T$. The vector $\boldsymbol{x}_T$ then allows to calculate $\hat{\boldsymbol{y}}_P$ as

$$\hat{\boldsymbol{y}}_P = \boldsymbol{C}_{PT}\boldsymbol{x}_T \,.$$

8.2.4 Reduction of the training data

Previous reports on ε-SVR and BLP hybrid prediction have assumed the availability of phenotypic measurements on a large number of hybrids. For both prediction methods, a reduction in prediction accuracy is to be expected if the size of the training set is decreased. A large sample size does, however, not necessarily imply a high prediction accuracy as the relevance of the training examples with respect to the future cross predictions, is of equal importance. Also the size and information content of the molecular fingerprints has an impact on the reliability of the prediction model as a smaller marker resolution implies a reduced chance of detecting marker-trait associations and less precise estimates of the genetic covariance between relatives.

Training sample size

In an attempt to assess the impact of the size of the training sample on the prediction accuracy of both ε-SVR and BLP, we employ three sampling schemes to obtain subsets of the original RAGT data set. For each sampling scheme, the prediction accuracy is determined in two ways: (1) by means of cross-validation on the training vectors $\hat{\boldsymbol{y}}_T^{rg}$ and $\hat{\boldsymbol{d}}_T^{rs}$ for predictions on total genotypic value and SCA respectively (2) by correlating against the validation vectors $\hat{\boldsymbol{y}}_V^{rg}$ and $\hat{\boldsymbol{d}}_V^{rs}$. The 38 hybrids that are common to training and validation data, are removed from the vectors $\hat{\boldsymbol{y}}_T^{rg}$ and $\hat{\boldsymbol{d}}_T^{rs}$ when the second prediction accuracy measure is used.

Random sampling For the random sampling scheme, the hybrids in the full training set are successively split at random to form smaller data sets from which ε-SVR and BLP prediction models are constructed. Initially, the prediction accuracy of both methods using all but one training examples is determined by means of a leave-one-out cross-validation (1). Predictions on the 49 hybrids that were tested in the validation field trial are obtained from ε-SVR and BLP models that were constructed from the 2316 non-validated hybrids (2). In the next step, the number of training examples made available to ε-SVR and BLP is cut in half and the cross-validation-based prediction accuracy is determined by making predictions on the other half of the training examples (1). The set of 2316 non-validated

hybrids is also randomly split in half and used to make ε-SVR and BLP-based predictions on the 49 validation hybrids (2). In subsequent steps, the number of training examples made available to ε-SVR and BLP is reduced further by randomly splitting the training data in 2^p pieces for $p = 1, \ldots, 6$. The whole process is repeated 100 times resulting in $100 \sum_{p=1}^{6} 2^p = 12600$ distinct ε-SVR and BLP prediction models.

Test-cross sampling The test-cross sampling scheme simulates the prediction of a hybrid formed by crossing a newly created inbred line with a well-known tester line. For each of the 197 inbred lines in the original data set, a separate ε-SVR and BLP prediction model is constructed using only information from hybrids that are not a child of that particular inbred. The resulting prediction models are used to predict the performance of the left-out hybrids and those hybrids in the validation data set that also have that particular inbred line as a parent. This sampling scheme therefore results in two predictions for each hybrid as both parental inbred lines function once as tester and once as newly developed line. In a balanced mating design (i.e. all 9960 distinct crosses between the Iodent and the ISSS lines are made), this sampling scheme would allow to assess the obtained prediction accuracy for Type 1 hybrids as defined by Schrag et al. (2009, 2010).

New-cross sampling The third sampling scheme simulates the prediction of a hybrid formed by crossing two newly developed inbred lines. Although this situation is rather uncommon in hybrid breeding programmes, it allows to compare ε-SVR and BLP in a worst-case scenario. For each hybrid in the dataset, a specific ε-SVR and BLP prediction model is constructed by removing all hybrids from the training set that have a parental inbred line in common with the selected hybrid. This sampling scheme relates to the Type 0 hybrids of Schrag et al. (2009, 2010).

Molecular marker information content

The impact of the information content of the molecular fingerprints is examined by taking random subsets of the available SSR or AFLP markers and subsequent construction of the ε-SVR and BLP prediction models. Again, prediction accuracy is determined by means of (1) cross-validation and (2) correlating against the estimates obtained from the validation trial. The size of the set of predictor markers is reduced in steps of 10% of the original fingerprint size and at each step, 100 iterations of the sampling routine are performed. Reducing the set of molecular markers often results in a singular coancestry matrix which prevents its inversion during the construction of a BLP prediction model. This situation

occurs if the marker-based estimate of the variance matrix of the training hybrids is rank deficient and therefore does not allow for a unique solution of the system of linear equations in Eq. (8.7). Any estimated variance matrix should be at least positive semi-definite as explained in Chapter 6 but in the present case, the marker-based estimate of the genetic covariance matrix $\boldsymbol{V}_T$ should be strictly positive definite as its Cholesky decomposition is used to make predictions on new hybrids. If the estimated covariance matrix, obtained from the reduced set of molecular markers, is singular, we obtain the minimum norm, least squares solution to Eq. (8.7). Other solutions might result in higher correlations but without relying on the validation data, there is no biological justification for preferring these solutions over the least squares solution.

8.3 Results

8.3.1 Unbalanced data handling

Three quarters of the hybrids in the validation field trial have measurements in the unbalanced training data set. These 38 hybrids therefore allow to identify the best way of obtaining a single hybrid score from unbalanced phenotypic data. Table 8.3 gives an overview of the observed correlations between the different types of hybrid scores and the genotypic estimates obtained from the validation field trial measurements. The latter were collected during one growing season at three locations in the South of France and as such, represent only a small part of the G×E space spanned by the training data. The correlations presented are therefore susceptible to environmental changes but should, however, allow for a relative comparison between the different data handling methods.

8.3.2 Reduction of the training data

Training sample size

Random sampling Figure 8.1 shows the prediction accuracy obtained by ε-SVR and BLP prediction models that were constructed by reducing the initial set of the training examples in the vectors $\hat{\boldsymbol{y}}_T^{rg}$ and $\hat{\boldsymbol{d}}_T^{rs}$. $p = 0$ indicates that a leave-one-out cross-validation is performed and predictions for the validation trial hybrids are obtained from ε-SVR and BLP models that are trained on the full vector $\hat{\boldsymbol{y}}_T^{rg}$ or $\hat{\boldsymbol{d}}_T^{rs}$, minus the entries of the 38 common hybrids. For each of the 100 iterations at $p = 1, \ldots, 6$, the training hybrids are randomly assigned to one of 2^p subsets and for each of these subsets, an ε-SVR and BLP prediction model is

Table 8.3: Squared Pearson correlation coefficients between the different types of training scores ($\hat{\boldsymbol{y}}_T^{rp}$, $\hat{\boldsymbol{y}}_T^{rg}$, $\hat{\boldsymbol{y}}_T^{fg}$) and SCA BLUPs ($\hat{\boldsymbol{d}}_T^{rs}$) obtained from the unbalanced phenotypic data set and the scores ($\hat{\boldsymbol{y}}_V^{rg}$) and SCA estimates ($\hat{\boldsymbol{d}}_V^{rs}$) obtained from measurements taken in the balanced validation field trial for the 38 common hybrids. For each trait, the combination of scores with the highest correlation is set in bold.

	score vector	validation data $\hat{\boldsymbol{y}}_V^{rg}$ / $\hat{\boldsymbol{d}}_V^{rs}$		
		yield	moist. cont.	flowering
training data	$\hat{\boldsymbol{y}}_T^{rp}$	0.04	0.61	0.43
	$\hat{\boldsymbol{y}}_T^{rg}$	**0.19**	**0.79**	**0.72**
	$\hat{\boldsymbol{y}}_T^{fg}$	0.05	0.59	0.43
	$\hat{\boldsymbol{d}}_T^{rs}$	0.03	0.15	0.17

constructed. These models are subsequently used to make predictions on (1) all hybrids that are not included in the training subset and (2) the 49 hybrids tested in the validation field trial. Despite the promising cross-validation results for SCA values, the observed correlations for the SCA predictions of the 49 validation hybrids, indicate that predicting SCA values by training on this set of unbalanced phenotypic data, is well beyond the capabilities of both ε-SVR and BLP.

Test-cross and new-cross sampling Table 8.4 gives an overview of the BLP and ε-SVR prediction accuracies when the training set is reduced in a non-random fashion to simulate predictions on hybrids for which one or both parental inbred lines are new and therefore untested. Squared Pearson correlation coefficients between the entries in vector $\hat{\boldsymbol{y}}_T^{rg}$ and their SSR or AFLP-based cross-validation predictions are presented for both sampling schemes as well as the squared correlations between the entries of the validation set vectors $\hat{\boldsymbol{y}}_V^{rg}$ and their predictions.

Molecular marker information content

The sensitivity of both ε-SVR and BLP to a reduction in the size of the molecular fingerprint is shown in Figure 8.2 by means of box and whisker plots. The set of SSR and AFLP markers is reduced in steps of 10%. For each step, a random subset of markers is selected and used to construct an ε-SVR and BLP prediction model by training on all entries of the vector $\hat{\boldsymbol{y}}_T^{rg}$ minus the 38 hybrids that are tested in the validation set. Prediction accuracy

Figure 8.1: ε-SVR and BLP prediction accuracies obtained by training on subsets of the vector of genotypic values $\hat{\boldsymbol{y}}_T^{rg}$ and the vector of SCA BLUPs $\hat{\boldsymbol{d}}_T^{rs}$. At $p = 0$, a leave-one-out cross-validation is performed on the training data and predictions on the 49 hybrids are made by training on all 2316 training hybrids. At $p = 1, \ldots, 6$, an ε-SVR and BLP prediction model are constructed from the 2^p subsets of the original vectors and AFLP or SSR predictor information. For each of these models, predictions are made for all training hybrids that are not in the particular subset and all 49 hybrids of the validation data set. This subset assignment procedure is replicated 100 times. Accuracy is expressed as the median of the squared Pearson correlation coefficient between the predictions for all hybrids and their corresponding entries in the training vectors $\hat{\boldsymbol{y}}_T^{rg}$, $\hat{\boldsymbol{d}}_T^{rs}$ (suffix cross), and the validation vectors $\hat{\boldsymbol{y}}_V^{rg}$ and $\hat{\boldsymbol{d}}_V^{rs}$ (suffix valid). The error bars indicate the 0.25 and 0.75 quantiles of each sampling distribution.

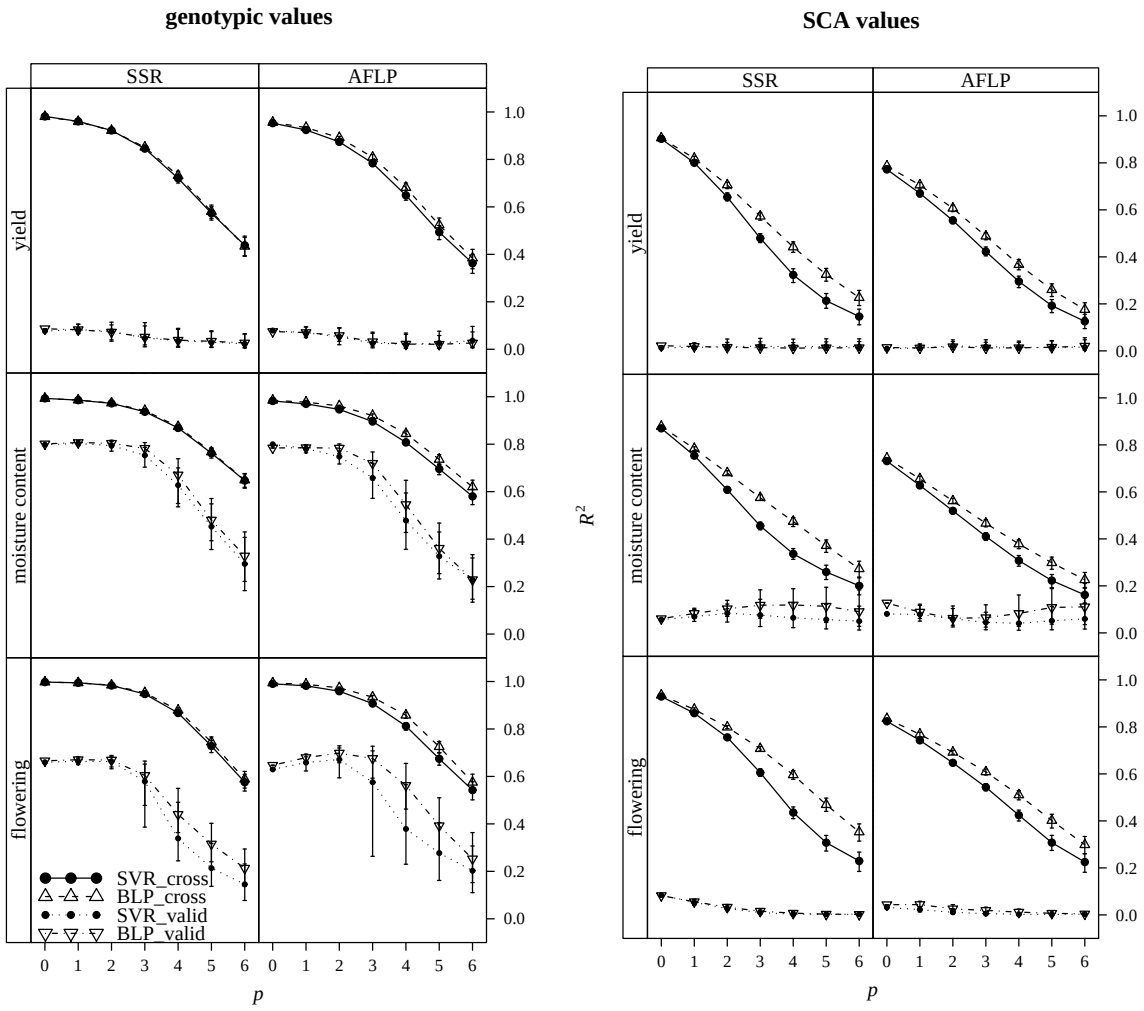


Table 8.4: Prediction accuracies, expressed as squared Pearson correlation coefficients, obtained from two sampling schemes simulating predictions on hybrids where one (Test-cross sampling) or both parents (New-cross sampling) are newly developed inbred lines. Cross-validation correlations on the vector $\hat{\boldsymbol{y}}_T^{rg}$ (cross) as well as correlations for predictions of the validation vector $\hat{\boldsymbol{y}}_V^{rg}$ (valid) are presented for the three traits grain yield, grain moisture content and days until flowering.

predictor	trait	predictand	Test-cross sampling		New-cross sampling	
			ε-SVR	BLP	ε-SVR	BLP
AFLP markers	yield	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.72	0.78	0.48	0.58
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.09	0.10	0.09	0.11
	moist.	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.80	0.85	0.63	0.71
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.53	0.67	0.31	0.58
	flower	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.80	0.84	0.62	0.69
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.30	0.43	0.04	0.22
SSR markers	yield	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.62	0.66	0.32	0.39
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.10	0.05	0.07	0.03
	moist.	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.77	0.72	0.57	0.51
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.41	0.38	0.15	0.14
	flower	cross ($\hat{\boldsymbol{y}}_T^{rg}$)	0.67	0.70	0.41	0.45
		valid ($\hat{\boldsymbol{y}}_V^{rg}$)	0.31	0.41	0.02	0.18

is expressed as squared Pearson correlation coefficients between the predictions of the 49 validation hybrids and their corresponding entries in the vector $\hat{\boldsymbol{y}}_V^{rg}$.

8.4 Discussion

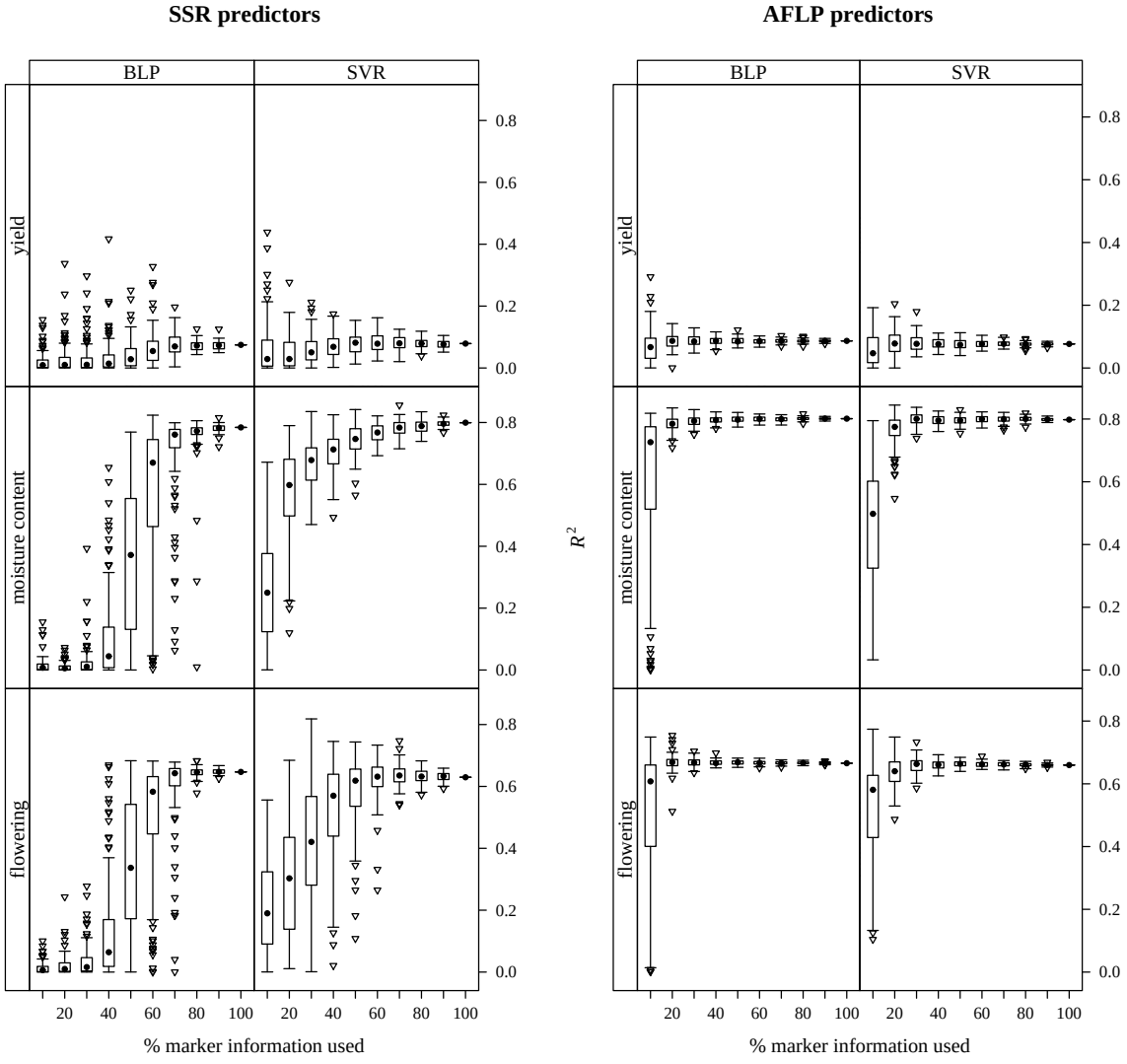
8.4.1 Unbalanced data handling

Predicting the phenotypic performance of untested hybrids by means of an ε-SVR or BLP model requires a training set of considerable size. Each training example should be represented by a single response value and a set of molecular marker-based predictors. A typical commercial hybrid breeding programme tests hundreds of new inbred combinations in a vast number of multi-location field trials on a yearly basis. The resulting data sets contain phenotypic measurements on numerous hybrids and would therefore allow for the construction of an ε-SVR or BLP prediction model at a low cost. However, the unbalanced nature of this kind of breeding data makes it hard to distill a single response value that allows to rank all hybrids on the same scale. We examined three mixed model-based methods to obtain such a score from unbalanced phenotypic data: (1) random phenotypes introduced by Bernardo (1994, 1995, 1996a,c), (2) random genotypes and (3) fixed genotypes. Besides these three types of genotypic scores, we also obtain an estimate of the SCA value for each hybrid in the training data.

The random genotypes approach results in the highest correlations for all three traits under study. The fixed genotypes approach seems to result in the lowest correlations and Bernardo's random phenotypes perform only slightly better. The inadequacy of the fixed genotypes is not unexpected because the assumption of fixed genotypic effects is likely to increase the standard error of the estimators of commercially uninteresting hybrids, as these have few records in the data set and no strength can be borrowed from records on related hybrids.The assumption of random nuisance effects on the other hand seems justified for this kind of breeding data as the number of levels of these factors is usually quite high.

Comparing the prediction accuracies of the three traits under study, we see that grain moisture content is the most promising trait for the construction of a reliable prediction model. The large contribution of the main genotypic effects (i.e. GCA and SCA) to the total variance (74%) and the low impact of the G×E components (14.3%) in the linear mixed model analysis with a random genotype assumption, explains these results. For the number of days until flowering, this partition is 44.5% versus 20% which results in

Figure 8.2: ε-SVR and BLP prediction accuracies obtained by constructing ε-SVR and BLP prediction models from the 2316 entries in vector $\hat{\boldsymbol{y}}_T^{rg}$ using subsets of the AFLP or SSR marker information as predictors for each of the three traits under study. Box and whisker plots show the range of squared Pearson correlation coefficients between the 49 entries in vector $\hat{\boldsymbol{y}}_V^{rg}$ and their predictions over 100 iterations of the marker sampling routine.



the somewhat lowered correlations observed for this trait. The trait grain yield, although of great interest to breeders, looks the least eligible candidate for the construction of a prediction model. This low correspondence between training and validation data estimates can be explained by the fact that the contribution of the G×E factors (38%) exceeds the contribution of the main genotypic factors (30.7%) to the total variance. The training examples are constructed marginal to the environmental factors such as growing season and geographical region while the validation data was collected at exactly one specific level of these factors. If a trait is subject to a large G×E variance, one can expect a genotypic effect, estimated over a large range of environments, to deviate substantially from an estimate obtained at one particular level of these environmental factors. A similar reasoning can explain the observed lack of correlation for the SCA effects although other aspects like the increased prediction error variance of the SCA BLUPs, the limited predictive value of a set of random SSR or AFLP markers with respect to a complex phenomenon like heterosis, and possibly reciprocal differences, also have their detrimental influence.

For the two promising traits moisture content and days until flowering the actual prediction accuracies obtained by ε-SVR and BLP models, when trained on the vectors of random genotypes, are quite close to the theoretical upper bounds presented in Table 8.3. This can be seen from the SVR_valid and BLP_valid lines in Figure 8.1 at $p = 0$. These specific points are obtained by correlating the ε-SVR and BLP predictions of the 49 validation hybrids with their random genotypic estimates in vector $\hat{\boldsymbol{y}}_V^{rg}$.

Our results indicate that the random genotypes approach is the best way to obtain a single genotypic score for each hybrid in the training data. By contrast, Bernardo (1994, 1995, 1996a,c) makes predictions on new hybrids by fitting the vector of random phenotypes $\hat{\boldsymbol{y}}_T^{rp}$ in Eq. (8.6). The entries in the resulting vector $\hat{\boldsymbol{y}}_P$ are *sensu stricto* not best linear predictions as the procedure does not take into account the covariance structure that originated from the measurement adjustments involving estimated fixed effects. This observation seems of minor importance as cross-validation results indicate a superior prediction accuracy compared to several other methods (Charcosset et al., 1998). However, a more straightforward approach would be to simply fit a number of additional parameters for the missing GCA and SCA components of the untested hybrids into the variance structure of the linear mixed model. As there are no phenotypic measurements linked to these effects, the additional columns in the random design matrix can all be set to zero. The estimated values for these additional effects are true best linear unbiased predictions or BLUPs and allow to reconstruct the predicted genotypic value of an untested hybrid by means of Eq. (8.2). The downside of this approach is that for each new prediction, the full set of mixed model equations needs to be solved. Moreover, an assessment of prediction

accuracy by means of cross-validation routines is not only computationally exhausting, but often just not sensible as leaving out the phenotypic measurements on one or more hybrids might divide the training data in two or more disconnected subsets. In this scenario, each of the disconnected subsets contains measurements on a different, non-overlapping set of hybrids which are tested in a different set of environments. Contrasts involving random genotypic effects of hybrids that belong to different, disconnected data subsets are usually estimable but do not conform to the usual interpretation as they rely on the implicit assumption that the genetic levels among the different environmental subsets are equal (Laloé, 1993). To avoid these pitfalls, a BLP prediction based on the random genotypic scores of the training hybrids is the next best option.

8.4.2 Reduction of the training data

Training sample size

In the previous section we indicated that using random genotypes to train our ε-SVR and BLP prediction models should result in superior prediction accuracies compared to the alternatives examined. For this reason, we continue to work with the random genotypes to evaluate the impact of the training sample size on the prediction accuracy of both ε-SVR and BLP.

Random sampling In Figure 8.1 we see that the behaviour of ε-SVR is quite similar to that of BLP when the size of the training set is reduced in a random fashion. For both methods, it is very clear that the cross-validation-based prediction accuracies consistently overestimate their validation trial counterparts. This is more explicit for the low heritability trait grain yield than for the traits moisture content and days until flowering. The observed disparity can be explained by the specific set of G×E effects that affect the validation data while the estimates derived from the training data are marginal to all environmental effects. If G×E effects explain a large portion of the observed variance for a trait, the observed heritability will be reduced correspondingly, as is the case for grain yield.
If we focus on the prediction of total genotypic value, the accuracy of ε-SVR and BLP shares a similar downward trend when the size training set is reduced, although ε-SVR usually performs slightly worse than BLP. The fall in prediction accuracy starts somewhere between $p = 2$ and $p = 3$, which is the equivalent of using 25% and 12.5% of the original training data respectively. If the training set is further reduced, the sampling variance of the validation trial-based prediction accuracies increases, as indicated by the widening of

the interquantile ranges. This increase in sampling error is less pronounced for the cross-validation-based prediction accuracies, giving a false indication of confidence for these favourable estimates. For the three traits under study, there is little difference between the behaviour of prediction models based on SSR markers and those using AFLP markers as predictors when the set of training hybrids is reduced by random selection.
If we focus on the prediction of SCA, we see that neither ε-SVR or BLP succeed in raising the median validation prediction accuracy, expressed as a squared Pearson correlation coefficient, above 0.13. Most striking is that the prediction accuracy estimates obtained through cross-validation give the impression that both ε-SVR and BLP are quite capable of making SCA predictions with a reasonable accuracy, especially if the full training set is used. The more pronounced impact of G×E effects on SCA measurements is again the most likely culprit here.

Test-cross and new-cross sampling If a non-random selection of training hybrids is performed, the superiority of the AFLP predictors becomes apparent, as can be seen from Table 8.4. In all but 2 scenarios, the prediction models based on AFLP markers have a greater prediction accuracy compared to those based on SSR markers. Table 8.4 again demonstrates the upward bias of the cross-validation-based prediction accuracy estimates. The ε-SVR prediction models are generally inferior to BLP when it comes to predicting the phenotypic performance of hybrids for which at least one of the parental inbred lines has no offspring in the training set. If both parents are unknown, neither ε-SVR nor BLP succeeds in making reliable predictions as the highest validation trial prediction accuracy is 0.58 for a BLP model trained on the trait grain moisture content using AFLP markers as predictors. The combination of a high heritability for grain moisture content and the more informative AFLP markers as predictors should allow this BLP model to be used for screening purposes.

Molecular marker information content

Reducing the set of predictors, by randomly selecting a subset of markers, has a negative effect on the prediction accuracy of both ε-SVR and BLP as can be deduced from Figure 8.2. The effect of the number of genotyped markers on the prediction accuracy appears to be subject to the law of diminishing marginal returns and little improvement is to be expected by further saturating the molecular fingerprint with additional AFLP or SSR markers. In this respect, Frisch et al. (2010) even observe a decline in prediction accuracy when the

number of genes for which expression data is incorporated in their transcriptome-based prediction models, is increased beyond a certain optimum.
The difference in behaviour between ε-SVR and BLP is most apparent for the traits grain moisture content and days until flowering in combination with the less informative SSR markers as predictors. As soon as 30% of the SSR markers are removed from the set of predictors, certain samples generate a substantially lower prediction accuracy of the BLP model while the accuracy of the equivalent ε-SVR model is nearly identical to that of the full marker set. Reducing the set of SSR predictors beyond this level, further inflates the sampling error of the BLP prediction accuracies, while at the same time the median of the distribution starts its steep descent. ε-SVR handles a reduction of the SSR predictors better than BLP as the sampling error starts to increase at lower values of the fingerprint size, while the median of the prediction accuracy shows a gentle decline as the number of predictors is reduced. The median ε-SVR prediction accuracy is for instance always superior to that of BLP as soon as 40% of the markers is removed. This observed superiority of ε-SVR over BLP is less pronounced if we use the AFLP markers as predictors. Both methods retain a good and comparable prediction accuracy for the traits grain moisture content and days until flowering, even when the set of AFLP predictors is reduced to 20% of its original size. Beyond this level, the prediction accuracy rapidly declines, while the sample variance increases. Even if 90% of the AFLP markers are removed, which is equivalent to a predictor set size of 57 dominant markers, several samples allow ε-SVR and BLP to obtain good prediction accuracies. Moreover, several samples of AFLP and SSR markers result in prediction accuracies that are greater than that of the equivalent model using the full set of markers. These observations indicate that an ε-SVR or BLP prediction model that uses only a specific subset of markers, might possibly improve the presented prediction accuracies but further study is needed to ascertain this point.

8.4.3 Conclusions

To construct an ε-SVR or BLP model for the prediction of phenotypic response based on a hybrid's molecular fingerprint, training data that contains a vector of marker scores and a single response value for every hybrid is needed. The best prediction accuracy is achieved by constructing these hybrid response values by summing the appropriate GCA and SCA BLUPs, obtained from a linear mixed model analysis with a random genotypic effects assumption.
If prediction accuracy is determined by means of a validation trial, both ε-SVR and BLP perform close to the theoretical limit for the traits grain moisture content and days until

flowering while they both fall short for grain yield, a trait with a low heritability in advanced breeding pools. The accuracy of SCA predictions is similarly insufficient for all three traits. This lack of predictive power is not reflected in the prediction accuracy measures obtained through cross-validation procedures, as these do not take into account the uncertainty introduced by G×E effects. Furthermore, if only a limited set of training examples is available but the genotyped markers are either numerous or very informative, BLP is more accurate than ε-SVR. If on the other hand the set of molecular markers is either restricted in size or information content, ε-SVR is the preferred prediction method.

CHAPTER 9

General conclusions and future prospects

The research presented in this dissertation was focused on the prediction of the agronomic performance of maize hybrids. The goal was to develop a methodology that would allow hybrid breeders to construct a prediction model using the phenotypic data that has resulted from earlier genetic evaluation trials. The proposed prediction models should use the molecular fingerprints of the parental inbred lines as explanatory variables. In Chapter 1, four research objectives were defined which are basically reformulations of the problems that were encountered while developing the ε-SVR-based prediction models. In this chapter, each of these research objectives is revisited and confronted with the conclusions reached in the different chapters of this dissertation. Current shortcomings and ideas for future improvements of the presented methods are discussed.

9.1 Data selection

The first research objective is related to the unbalanced nature of the genetic evaluation data that is generated in commercial breeding programmes. The higher the level of unbalancedness, the lower the actual information content of the available set of phenotypic measurements which in turn reduces the accuracy of the resulting prediction model. It is a fair assumption that the total genotyping budget for the construction of a hybrid prediction model will always be restricted in some way. This assumption implies that the accuracy of the final prediction model depends on finding the optimal trade-off between the number of genotyped inbred lines and the density of their molecular fingerprint. However, finding this optimal number of inbred lines is somewhat pointless if one cannot identify the most informative set of inbred lines of that particular size. In a similar reasoning, the deduction of the optimal number of molecular markers is inherently associated with the problem of

finding the subset of markers which has a maximal genome coverage. The first research objective therefore aims at finding a solution for these highly related problems which is basically the topic of Chapter 5.

In this chapter, algorithms from the field of graph theory are described which allow to solve the described optimisation problems. The problem of finding the maximally informative subset of unbalanced phenotypic data is handled by iteratively solving the 'discrete p-dispersion problem' by means of a means of a maximum clique-based algorithm. The described procedure makes use of the generalised coefficient of determination (CD) of pairwise contrasts which can be obtained as a by-product of a linear mixed model analysis of the available phenotypic data. It is shown that this approach allows to select a fixed set of hybrids for which the available phenotypic measurements are both highly replicated and balanced. It should, however, be clear that the described selection procedure is still open to improvement. For instance, the CD of pairwise contrasts is strongly associated with the quality of the experimental design (Bueno and Gilmour, 2003), but this does not necessarily imply that the data subset with the highest minimum CD of pairwise contrasts is also identified as the maximally informative subset according to any of the other documented measures of experimental design quality. The CD itself, as a measure of contrast quality, has several flaws. If two genotypes are disconnected, the CD of their pairwise contrast does not necessarily equal zero. This shortcoming can be compensated by post-correcting the CD of disconnected genotypes but this obviously requires the identification of all disconnected subsets which can be computationally expensive. Furthermore, the CD is claimed to make a trade-off between data quantity (i.e. replication) and data quality (i.e. balance) (Laloé et al., 1996) but the actual properties (e.g. monotonicity) of this trade-off principle are unclear and require further study.

However, the biggest problem of the described procedure is the bias towards good performing genotypes. Newly developed hybrids which demonstrate a competitive level of agronomic performance will be evaluated many times more than bad performing genotypes. This means that contrasts involving good performing hybrids will generally have high CD values which makes them more attractive for selection. As a result, the final prediction model is likely to be positively biased as the majority of the training examples will have a good track record. The only cure for this disease is to abandon the CD and develop a new selection criterion that tries to maximise the variance of the genotypic BLUPs whilst ensuring that both the data quality and data quantity are taken into account. It should be clear that the construction of such a criterion that needs to find a balance between these three competing objectives, is a non-trivial task. Furthermore, the described optimisation procedure, based on the discrete p-dispersion graph theory problem, requires this criterion

to quantify the quality of each individual pairwise contrast. This requirement will make it very hard, if not impossible, to introduce some measure of total BLUP variance in the selection criterion.

The CD-based data selection procedure selects hybrids, while the genotyping budget puts constraints on the number of inbred lines. Obviously, selecting hybrids based on data information content indirectly selects parental inbred lines but this procedure does not necessarily maximise the number of training examples. This problem setting was translated to the k-densest subgraph problem from graph theory for which several efficient heuristics are available. The data quality and quantity constraints are enforced by a CD-based preselection of hybrids. However, the results described in Chapter 5 indicate that the avoidance of disconnected genotypes is the only important issue. It might therefore be worthwile to identify the most productive parents (in terms of progeny size) first, and then subsequently remove any disconnected hybrids in this selection. Although this might seem like a promising strategy to explore, this approach does not provide an escape from the non-optimal nature of sequentially performing two separate selection procedures. The integration of both selection procedures in a single, multi-objective optimisation problem is a topic which will require extensive study. Furthermore, it might not even be that beneficial to maximise the number of progeny from a fixed set of parents as this might handicap the generalising capabilities of the final prediction model due to the relatedness of the training examples. Again, further study is needed to clarify these issues.

9.2 Marker-based coancestry estimation

The second research objective entails the development of a marker-based coancestry estimation procedure that is PSD, always produces estimates within the unit interval and is specifically designed for use in hybrid breeding programmes. The Weighted Alikeness in State or WAIS estimator, discussed in Chapter 6, meets all these requirements. The behaviour of WAIS is compared to that of other CoC estimators under a typical hybrid breeding selection scheme by means of simulations and the RAGT R2n maize breeding data. It is clear that WAIS can compete with other popular CoC estimators although it generally does not take first prize when it comes to producing a superior linear mixed model fit or a minimal mean squared error. This is most likely an unfortunate result of constraining the CoC estimator to be PSD and comparing it against a set of unconstrained alternatives. The WAIS estimator is nevertheless a valuable asset for hybrid breeders as it allows for the carefree modelling of the covariance between hybrids in linear mixed

model settings encountered in breeding value estimation, association studies and genomic selection. Besides this WAIS estimator, Chapter 6 also introduces an MCMC-based matrix bending procedure, which allows to transform non-PSD CoC matrices to their nearest PSD neighbour which can be used to model the covariance between specific genetic components in a linear mixed model analysis. The examined CoC estimators and matrix bending procedures are made available in the form of a free software package named CoCoa.

The use of the WAIS estimator is restricted to hybrid breeding programmes as the estimation of allelic weights requires an unrelated reference set of genotypes. Expanding the target audience therefore requires a modification of the WAIS formulation to allow for mixed breeding pools. Two strategies towards handling these mixed breeding pools have in fact already been explored. The first strategy assumes the availability of accurate and detailed pedigree information which allows to identify unrelated genotypes which in turn allow for the estimation of the required allelic weights. This approach is obviously only valid if there are many unrelated genotypes in the breeding pool, which is generally not the case. Furthermore, if pedigree information is available, one might be better off by simply calculating the pedigree-based CoC estimator as it is shown in Chapter 6 to perform better than the examined marker-based alternatives. The second strategy assumes that the origin of each allele in a mixed breeding pool can be reconstructed by means of a Bayesian algorithm like implemented in the popular software package "structure" (Pritchard et al., 2000). Reformulation of the WAIS estimator to incorporate these estimates of allelic origin was rather straightforward but the actual performance of the resulting CoC estimator was quite disappointing. By means of simulations, it was discovered that these gloomy results were not caused by the reformulation of the WAIS estimator but were due to the imprecision of the structure-based population origin estimates. This is not unexpected as the Bayesian algorithm used for inference of the population structure relies heavily on the assumption of Hardy-Weinberg equilibrium, which is generally subject to gross violations in commercial breeding pools. These two reformulations of WAIS, corresponding to the two described expansion strategies are not presented in this dissertation.

An interesting but still unexplored strategy to expand the applicability of the WAIS estimator is inspired by the work presented by Stich et al. (2008). The idea is to incorporate the WAIS formulation (i.e. Eq. (6.7)) directly in the variance structure of the linear mixed model that is being fitted on the available phenotypic data. The required weights in matrix $\boldsymbol{W}$ can, for example, be parametrised to originate from a zero centred normal distribution for which the variance can be estimated by means of (restricted) maximum likelihood. This approach would effectively allow to use the WAIS CoC estimator in mixed breeding pools without the need for pedigree data.

9.3 ε-SVR for genomic prediction

In Chapter 7, the selected subset of phenotypic data is analysed by means of a linear mixed model. The levels of the nuisance factors like METs, locations and blocks are modelled as fixed effects and these estimates are subsequently used to correct each of the phenotypic measurements to obtain what are called 'random phenotypes' in Chapter 8. These random phenotypes serve as training examples for the construction of ε-SVR prediction models for the traits yield, grain moisture content and days until flowering. By means of a leave-one-out cross-validation strategy it was established that a kernel-induced non-linear ε-SVR function generally provides a better prediction accuracy than its linear counterpart. The genetic distance-based kernel functions can generally compete with the popular Gaussian kernel. More astonishing however, is the competitive performance of the BLP approach, an intrinsically linear function which does not require a computationally expensive grid-search for finding appropriate values for model parameters. At this point, the only real advantage that ε-SVR has to offer over BLP is its flexibility towards combining different types of molecular markers as predictors.

The reported prediction accuracies in Chapter 7 leave much room for improvement. The idea of performing a feature selection by means of the Recursive Feature Elimination strategy (Guyon et al., 2002), specifically adapted and implemented for the regression framework, turned out to provide little or no consolation. In a similar way, the development of problem-specific kernel functions did not bring about large improvements in the reported prediction accuracies. These observations fastened the suspicion on the preprocessing step, more specifically the linear mixed model specification. The observant reader might have noticed that the linear mixed models used in Chapter 6 are more elaborate than the model used in Chapter 7 which, for example, does not fit any G×E effects. The reason for this deliberate simplification of the preprocessing step is purely technical. The various linear mixed model analyses described in Chapter 7 were performed on a computer with a limited memory size (1 Gb to be precise). This memory limitation restricts the number of mixed model equations that ASReml can handle which basically means that only simplified models can be fitted to a data set of such size. A year later, it was decided to buy a dedicated 64-bit workstation having a total memory size of 32 Gb which allowed to fit more appropriate mixed models, incorporating additional fixed nuisance effects and various G×E terms. Furthermore, non-essential check varieties that do not connect the different METs were dropped which reduced the number of checks from 33991 to 3022. The research presented in Chapter 7 was performed before that of Chapter 6, explaining the apparent discrepancy

in the various mixed model specifications.

In Chapter 8, the data preprocessing step was examined more thoroughly. Several strategies based on an appropriate linear mixed model analysis were explored, allowing to conclude that the 'random genotypes' approach results in superior estimates of the agronomic performance of the maize hybrids. The improvement in leave-one-out cross-validation-based prediction accuracy of both ε-SVR and BLP is astonishing, approaching R^2 values of 0.99 for all traits and marker types. In fact, these cross-validation-based prediction accuracies are so good that even the most gullible reader should start questioning their reliability. To examine the accuracy of these models without relying on cross-validation, a specific field trial was designed measuring the agronomic performance of 49 inter-heterotic hybrids in three locations in the South of France. The results were less favourable but still optimistic for the traits moisture content and days until flowering. Unfortunately, the obtained prediction accuracy for grain yield, the most important trait from a commercial perspective, left little hope for the routine application of ε-SVR or BLP in a commercial breeding programme. A similar conclusion was reached for all examined traits concerning the prediction of the specific combining ability, an estimator for the heterosis effect.

The validation field trial-based prediction accuracies differ substantially between the examined traits. This can be explained by noticing that the linear mixed model analysis of the RAGT R2n data predicts hybrid performance marginal to the effects of growing seasons, locations, METs,.... The analysis of the validation field trial on the other hand, predicts conditional on one specific year and geographical region. It is therefore well within expectations that these two type of predictions differ, where the level of discrepancy depends on the relative impact of the different G×E components on the total phenotypic variance. Compared to grain yield, the traits grain moisture content and days until flowering have a reasonably high heritability which can be attributed at least partially to the small contribution of the G×E components to the variance. For grain yield, the high impact of G×E variance components explains why the marginal and conditional predictors differ so much. In this situation, both the ε-SVR and BLP prediction models are constructed from training examples that deviate considerably from their validation trial equivalents, evidently resulting in a low estimates for prediction accuracy.

An interesting idea is therefore to construct ε-SVR or BLP prediction models that are conditional on a specific geographical region. The methods presented by Welham et al. (2004) allow to obtain training examples that are conditional on a particular level of one or more factors. These conditional predictors of hybrid performance allow to construct ε-SVR and BLP prediction models that are specific for that particular environment. Although these environment-dependent prediction models are likely to demonstrate an improved

prediction accuracy, this idea has not been pursued further. The biggest hurdle is the definition of an environment. We can base this definition on the proximity of the field trials in time and/or space or the similarity of their phenotypic measurements or environmental conditions. The exploration of these possibilities is a promising research topic on its own, worthy of the undivided attention that can be provided by a 'fresh' Ph.D. study.

In an attempt to identify both the strengths and weaknesses of ε-SVR and BLP, several data reduction scenarios were also examined in Chapter 8. The conclusions are reasonably straightforward, BLP performs best when the inbred lines are genotyped with large and informative molecular fingerprints and is not very sensitive to a reduction in the number of training examples. ε-SVR on the other hand, requires a large number of training examples but performs well even if the molecular fingerprints are small or not very informative. This conclusion confirms the findings on optimal data selection strategies in Chapter 5.

It is to be expected that fingerprinting the inbred lines by means of a large number of SNP markers will not substantially improve the reported prediction accuracies of ε-SVR or even BLP. This statement might be somewhat counterintuitive as it is generally assumed that saturating the genetic map facilitates the detection of marker-trait associations. However, Remington et al. (2001) demonstrate that in maize, LD between SNPs declines very rapidly with distance while at the same time they provide strong evidence of substantial genomewide LD between a set of SSR markers. Stich et al. (2006) also find extensive LD blocks using both SSR and AFLP markers in a collection of elite maize inbreds while a more recent study of Yan et al. (2009) confirms the fast LD decay of SNP markers in a highly diverse global maize collection. Remington et al. (2001) explain these contrasting results by suggesting that most SNP alleles predate the domestication of maize while the highly variable SSR alleles, being much more sensitive to mutation, have predominantly arisen during the domestication process. Despite the rather speculative nature of this claim, it should be clear that further research is required to assess the added value of high density SNP arrays with respect to the prediction of hybrid maize performance.

The results of the random marker reduction procedures do, however, indicate that the development of more intelligent feature reduction strategies might allow to improve the reported prediction accuracies of both ε-SVR and BLP. This research opportunity leans towards the currently very active topic of genomewide association studies and as such, might fill the gap between two of today's most auspicious directions of quantitative genetic research.

9.4 Conclusions

The methodology presented in this dissertation allows to predict the agronomic performance of possibly non-existing maize hybrids from the phenotypic measurements that were collected on other hybrids in past evaluation trials. The approach consists of two distinct steps: first, a linear mixed model analysis is performed to obtain training examples from the unbalanced phenotypic data and in the second step, these training examples are used to construct a prediction model by means of ε-SVR. This two-step approach is obviously suboptimal and in this respect, the direct integration of a RKHS regression function in a linear mixed model framework is a very exciting development that has recently been published by Gianola and van Kaam (2008). Although RKHS regression does not provide the sparsity of an ε-SVR function, it should be clear that is a particularly promising direction of research, which might allow to improve the prediction accuracies reported in this dissertation. The further development and exploration of this unified framework will very likely turn out to be a long, but nevertheless interesting journey, filled with the numerous detours, roundabouts and dead-ends that lie at the heart of true scientific endeavour.

Summary

Genomic selection is a breeding strategy in which superior genotypes are identified based on a direct analysis of their molecular fingerprints, therefore making phenotypic records redundant. This is a promising and very active topic of research as genotyping costs are steadily decreasing and next-generation sequencing technology holds the promise of making complete sequence information available at a reasonable cost. At the heart of every genomic selection approach lies a genomic prediction model. This genomic prediction model is provided with a molecular fingerprint of one or several candidate genotypes and produces some form of output which allows the breeder to identify the most promising genotypes. Such marker-based prediction models are particularly useful in hybrid breeding programmes as the homozygosity of the parents allows to assess the agronomic performance of their offspring before these are even conceived. A reliable genomic prediction model is therefore expected to have a considerable impact on the cost-effectiveness of hybrid breeding programmes. This is particularly the case for maize breeding programmes where the development of in-vivo haploid inducer lines allows to obtain fully homozygous inbred lines in a single generation.

If the trait under study is regulated by a small number of genes, admitting to a Mendelian inheritance pattern, the genotyping of a limited number of associated markers is generally sufficient to identify the desired genotypes without additional phenotyping efforts. Unfortunately, the traits with agronomic importance are generally of a more quantitative nature, exhibiting a continuous distribution of phenotypic values over the set of candidate genotypes. The prediction of these traits requires some form of regression modelling. The classical approaches to linear and non-linear regression are however not suited for handling the large number of predictors that are provided by dense molecular fingerprints and stepwise model selection techniques suffer from inflated family-wise type I errors.

The basic objective of this dissertation is to explore the capabilities of kernel-based techniques, more specifically ε-insensitive support vector machine regression (ε-SVR), for genomic prediction in hybrid maize. The presented research uses the phenotypic and molecu-

lar information that was routinely generated between 1984 and 2005 as part of the the grain maize breeding programme of the private company RAGT R2n. Using available breeding data has the obvious advantage of allowing for a low-budget research, but also ensures that the resulting prediction models can effectively be implemented on top of existing commercial or non-commercial (hybrid) breeding programmes.

The available breeding data is unfortunately very unbalanced. Some parental lines are reused many times in different inbred line combinations while others only appear once in the company's pedigree. Furthermore, some hybrids are tested many times under various environmental conditions while others only have a single phenotypic record. For these reasons, the available phenotypic data is preprocessed by means of a linear mixed model analysis which allows to obtain a single phenotypic score for every hybrid. These scores can be used to construct an ε-SVR-based prediction model. Besides being unbalanced, the RAGT R2n data set is also very large, containing records on thousands of inbred lines and hybrids. As the genotyping budget is limited, only a restricted set of inbred lines can be included in the study. Given a fixed genotyping budget, one still needs to find the optimal trade-off between the number of genotyped inbred lines and the density of their molecular fingerprint, where optimal refers to the prediction accuracy of the resulting ε-SVR-based prediction model.

In this dissertation it is shown, by means of a simulation study, how the optimal number of inbred lines and molecular markers can be determined when confronted with a fixed genotyping budget, a genetic map and a set of unbalanced phenotypic measurements on hybrid genotypes. It is demonstrated how efficient algorithms for solving the 'discrete p-dispersion problem' from the field of graph theory, allow to select the set of hybrids with predefined cardinality that has the most informative phenotypic measurements, where informative refers to both the level of balance and level of replication. It is also shown how to select a fixed-size subset of molecular markers with maximal genome coverage, a problem that can be translated to a one-dimensional variant of the discrete p-dispersion problem. Efficient algorithms that solve the 'densest k-subgraph problem' from graph theory, allow to maximise the number of training examples by selecting parental inbred lines that have produced the maximum number of offspring amongst themselves. The combination of these graph-based algorithms, solving different types of selection problems, allows to identify the most promising data subset for the construction of an ε-SVR-based genomic prediction model.

A linear mixed model analysis allows to obtain a single phenotypic score for every hybrid in an unbalanced (sub)set of phenotypic measurements. In these models, nuisance factors like trial, location and blocks are generally modelled as fixed effects while genetic components

like GCA and SCA values are fitted as random variables. The covariance of these random variables is often assumed to be a function of the coefficients of coancestry (CoC) between pairs of inbred lines. The CoC between two inbred lines can be estimated from detailed and accurate information concerning their pedigree backgrounds. However, if the selection history of these inbred lines is no longer available or has become too complex for a classical pedigree analysis, CoC estimates can also be obtained from their molecular fingerprints. The field of population genetics has several CoC estimation procedures at its disposal, but if the genotyped individuals are highly selected inbred lines, their application is not warranted as the theoretical assumptions on which these estimators were built, usually linkage equilibrium between marker loci or even Hardy-Weinberg equilibrium, are not met. An alternative approach requires the availability of a genotyped reference set of inbred lines, which allows to correct the observed marker similarities for their inherent upward bias when used as a coancestry measure. However, this approach does not guarantee that the resulting coancestry matrix is at least positive semi-definite (PSD), a necessary condition for its use as a covariance matrix.

In this dissertation, a new CoC estimator named the weighted alikeness in state or WAIS estimator is presented. This marker-based coancestry estimator is compared to several other commonly applied relatedness estimators under realistic hybrid breeding conditions in a number of simulations. We also fit a linear mixed model to the RAGT R2n data and compare the likelihood of the different variance structures. WAIS is shown to be PSD, which makes it suitable for modelling the covariance between genetic components in linear mixed models involved in breeding value estimation or association studies. Results indicate that it generally produces a low root mean squared error under different breeding circumstances and provides a fit to the data that is comparable to that of several other marker-based alternatives. Recommendations for each of the examined coancestry measures are provided.

Fitting the WAIS estimator in the variance structure of a linear mixed model does not always produce the best model fit. Sometimes, another CoC estimator might fit better to the available data, despite it not being a PSD measure. In this case, a matrix bending routine can be used to bend the matrix towards its nearest PSD equivalent. In this dissertation, a new MCMC-based bending procedure is presented and compared with a more classic bending procedure based on a singular value decomposition. WAIS and four other CoC estimation procedures, the two examined matrix bending routines and several other matrix manipulation tools were implemented in the software package CoCoa, which is made freely available under the conditions of the GNU General Public License.

Initially, a very basic linear mixed model is fitted to the selected subset of phenotypic

measurements and estimates for the effects of each of the fitted nuisance effects are obtained. These estimates allow to correct each of the phenotypic measurements and to obtain a single score, denoted as a 'random phenotype' for every hybrid. These random phenotypes are subsequently used to construct genomic prediction models based on ε-SVR and Best Linear Prediction (BLP). The use of kernel functions allows ε-SVR to fit a linear model in a high-dimensional feature space, which becomes a non-linear model in the original input space. The performance of all-rounder kernel functions like the Gaussian kernel are examined and it is shown how PSD CoC measures can also be used as kernel functions. Initial results using the RAGT R2n data, indicate that ε-SVR and BLP match each other's prediction accuracies for several combinations of marker types and traits. The ε-SVR framework, however, allows for a greater flexibility in combining different kinds of predictor variables. At this stage, the reported cross-validation-based accuracy measures for the prediction of grain yield are insufficient to allow for an efficient genomic selection. In an attempt to improve the prediction accuracy, the preprocessing step by means of a linear mixed model analysis is further examined. Additional fixed nuisance factors and G×E terms are introduced in the model formulation, resulting in more appropriate models for the mean and the variance. It is shown that a simple summation of BLUPs provides much better training examples for the construction of ε-SVR or BLP prediction models, compared to the random phenotypes approach. These modifications result in cross-validation-based prediction accuracy measures that are extremely good, even when predicting SCA values. Therefore, the results of a specifically designed validation field trial, which consist of testing 49 hybrids in three locations in the South of France, are to give a definitive judgement on the predictive capabilities of both ε-SVR and BLP.

These results indicate a considerable discrepancy between prediction accuracies obtained by cross-validation procedures and those obtained by correlating the predictions with the results of the validation field trial. The reason for this is fairly obvious, the training examples are predicted marginal to the effects of growing seasons and locations, while the validation hybrids are predicted conditional on specific levels of these factors. The amount of discrepancy depends on the trait under study. For grain yield, the reported correlations between predicted and measured phenotypic values leave little hope for a reliable genomic selection. This trait has a low heritability in advanced breeding pools, which is mainly a result of the high contribution of G×E effects to the total phenotypic variance which in turn is responsible for the poor correlations between marginal and conditional predictions. The other two examined traits, namely grain moisture content and days until flowering, have a higher heritability and are therefore predicted more accurately by ε-SVR and BLP. The limits of the predictive capabilities of these two methods are further examined by

reducing the number of training hybrids and the size of the molecular fingerprints. The prediction accuracy of BLP turns out to be less sensitive to a reduction of the number of training examples compared to that of ε-SVR. The latter is, however, better at predicting hybrid performance when the size of the molecular fingerprints is reduced, especially if the initial set of markers has a low information content.

Samenvatting

Genomische selectie is een veredelingsstrategie waarbij de superieure genotypes geïdentificeerd worden door een rechtstreekse analyse van hun moleculaire vingerafdrukken, hetgeen fenotypische waarnemingen overbodig zou moeten maken. Dit is een zeer actief en veelbelovend onderzoeksdomein, onder meer als gevolg van de dalende kostprijs van moleculaire merkers en de recente ontwikkelingen op het gebied van 'next-generation sequencing'. Deze laatste zouden het in de nabije toekomst mogelijk moeten maken om de volledige DNA sequentie te bepalen van de kandidaat-genotypes tegen een economisch realistische prijs. De kern van elke genomische selectie is de ontwikkeling van een genomisch predictiemodel. Een dergelijk predictiemodel moet, aan de hand van een moleculaire vingerafdruk, een quotatie opleveren waardoor de veredelaar de meest beloftevolle genotypes kan identificeren. Deze, op moleculaire merkers gebaseerde predictiemodellen, zijn vooral nuttig in hybride veredelingsprogramma's aangezien de homozygotie van de ouders toelaat om de agronomische prestaties van hun nakomelingen in te schatten alvorens deze verwekt worden. Een betrouwbaar genomisch predictiemodel zal ongetwijfeld een grote impact hebben op het rendement van hybride veredelingsprogramma's. Dit is zeker het geval bij maïs, waar de ontwikkeling van specifieke lijnen die in-vivo haploïdie induceren, het toelaat om in één enkele generatie volledig homozygote inteeltlijnen te bekomen.

Als het te voorspellen kenmerk gereguleerd wordt door één of enkele genen die Mendeliaans overerven, dan is het meestal mogelijk om de gewenste genotypes te identificeren aan de hand van een beperkt aantal moleculaire merkers die geassocieerd zijn met dit kenmerk. Helaas zijn de agronomisch belangrijke kenmerken eerder van kwantitatieve aard, waardoor er een continue verdeling van de fenotypische waarnemingen over de kandidaat-genotypes wordt waargenomen. De voorspelling van dit soort kenmerken vereist een vorm van regressie. De klassieke methodes voor lineaire en niet-lineaire regressie zijn echter niet geschikt voor de verwerking van het grote aantal predictoren die een uitgebreide moleculaire vingerafdruk beschikbaar maakt. Stapsgewijze modelleertechnieken hebben dan weer te lijden onder een verlaagde specificiteit die wordt veroorzaakt door het grote aantal statistische

toetsen dat noodzakelijk is.
Het belangrijkste doel van deze doctoraatsthesis is dan ook het verkennen van kernel-gebaseerde methodes, meer specifiek ε-insensitive support vector machine regression (ε-SVR), voor het ondersteunen van genomische selectie bij maïshybriden. Het voorgestelde onderzoek maakt gebruik van gegevens die gegenereerd werden tussen 1984 en 2005 als onderdeel van het korrelmaïsveredelingsprogramma van de private onderneming RAGT R2n. Het gebruik van deze kant-en-klare fenotypische gegevens heeft het voordeel dat dit onderzoek met een zeer bescheiden budget kan worden uitgevoerd en verzekert bovendien dat de resulterende predictiemodellen kunnen geïmplementeerd worden in commerciële en niet-commerciële (hybride) veredelingsprogramma's.
De beschikbare veredelingsgegevens zijn echter weinig gebalanceerd. Bepaalde ouderlijnen worden herhaaldelijk gebruikt in verschillende kruisingen, terwijl andere slechts éénmalig voorkomen in de stamboom van het veredelingbedrijf. Bovendien worden sommige hybriden veelvuldig getest onder variërende omgevingsomstandigheden terwijl andere hybriden slechts één enkele fenotypische meting hebben. Om deze redenen wordt de beschikbare fenotypische informatie eerst geanalyseerd aan de hand van een gemengd lineair model dat toelaat om een enkele fenotypische score te berekenen voor elke hybride. Vervolgens kunnen deze scores gebruikt worden bij het opstellen van een ε-SVR-gebaseerd predictiemodel.
De beschikbare gegevens van RAGT R2n zijn naast ongebalanceerd ook zeer omvangrijk aangezien het gaat over de metingen van duizenden inteeltlijnen en hybriden. Gegeven het beperkte budget voor genotypering, wordt er dus een selectie gemaakt van de inteeltlijnen die opgenomen worden in deze studie. Bovendien moet men bij een vast genotyperingsbudget de optimale afweging kunnen maken tussen het aantal genotypes dat gescoord zal worden en de dichtheid van hun moleculaire vingerafdruk, waarbij het de bedoeling is om de accuraatheid van het resulterende ε-SVR-gebaseerde predictiemodel te maximaliseren.
In deze doctoraatsthesis wordt aan de hand van een simulatiestudie aangetoond hoe het optimaal aantal inteeltlijnen en moleculaire merkers kan bepaald worden, wanneer men geconfronteerd wordt met een vast genotyperingsbudget, een genetische kaart en een set ongebalanceerde fenotypische metingen van hybride genotypes. Deze studie toont aan hoe efficiënte algoritmes voor het oplossen van het 'discrete p-dispersie probleem' uit het gebied van de grafentheorie, kunnen aangewend worden om de meest informatieve subset van hybriden met vooraf gedefinieerde kardinaliteit te identificeren. Het informatief zijn slaat hier zowel op de gebalanceerdheid van de metingen alsook het aantal metingen zelf. Er wordt ook aangetoond hoe men op een genetische kaart een vast aantal moleculaire merkers kan selecteren zodat deze het genoom maximaal bedekken. Dit selectieprobleeem kan namelijk vertaald worden naar een eendimensionale versie van het discrete p-dispersie

probleem. Het aantal trainingsvoorbeelden kan men maximaliseren door het identificeren van de set van inteeltlijnen die onderling het meeste nakomelingen hebben geproduceerd en dit door gebruik te maken van efficiënte algoritmes die specifiek ontwikkeld werden voor het 'dichtste k-subgraaf probleem'. De combinatie van deze algoritmes uit de grafentheorie, die verschillende types van selectieproblemen oplossen, maakt het mogelijk om de meest beloftevolle data subset te identificeren voor het opstellen van een genomisch predictiemodel.

Een analyse aan de hand van een gemengd lineair model laat toe om één enkele fenotypische score te bekomen voor elke hybride in een ongebalanceerde dataset met fenotypische waarnemingen. In deze modellen worden omgevingsfactoren zoals proef, locatie en replicatie meestal als vaste effecten gemodelleerd, terwijl de genetische componenten als willekeurige variabelen worden ingepast. De covariantie van deze willekeurige variabelen wordt meestal verondersteld een functie van de verwantschapscoëfficiënt (CoC) tussen paren van inteeltlijnen te zijn. Men kan deze CoC schatten aan de hand van gedetailleerde en accurate informatie over de stamboom van de betrokken inteeltlijnen. Indien deze stamboominformatie niet meer beschikbaar is of dermate complex wordt dat een klassieke stamboomanalyse niet meer tot de mogelijkheden behoort, dan kunnen CoC-schatters ook verkregen worden door gebruik te maken van de moleculaire vingerafdrukken van de inteeltlijnen. In de populatiegenetica heeft men verschillende merkergebaseerde procedures om de CoC te schatten, maar deze lijken weinig geschikt als de gegenotypeerde individuen sterk geselecteerd en ingeteeld zijn. In dergelijk geval zijn de theoretische veronderstellingen waarop deze procedures steunen, meestal linkage equilibrium of zelfs Hardy-Weinberg evenwicht, namelijk helemaal niet voldaan. Een alternatieve aanpak veronderstelt de beschikbaarheid van een referentieset van onverwante, gegenotypeerde inteeltlijnen. Een dergelijke set laat toe om de gemiddelde merkersimilariteit tussen twee inteeltlijnen te corrigeren, aangezien dit anders een vertekende schatter voor verwantschap vormt. Deze aanpak garandeert echter niet dat de resulterende verwantschapsmatrix op zijn minst positief semi-definiet (PSD) zal zijn, een noodzakelijke voorwaarde als deze gebruikt wordt als een variantiematrix in een gemend lineair model.

In deze thesis wordt dan ook een nieuwe CoC-schatter voorgesteld, namelijk de 'Weighted Alikeness In State' of WAIS-schatter. Deze merkergebaseerde verwantschapsschatter wordt vergeleken met verschillende andere schatters door gebruik te maken van simulaties die de omstandigheden nabootsen die zich voordoen in een hybride veredelingsprogramma. De geselecteerde gegevens van RAGT R2n worden onderworpen aan een analyse met een gemengd lineair model waarbij de verschillende verwantschapsschatters in het model van de variantie worden ingepast. Dit laat toe de verschillende schatters te rangschikken op

basis van de probabiliteit van het resulterende lineaire model. Er wordt aangetoond dat WAIS altijd een PSD verwantschapsmatrix oplevert, hetgeen toelaat om deze schatter te gebruiken om de covariantie tussen genetische componenten te modelleren in gemengde lineaire modellen voor het schatten van kweekwaardes of associatiestudies. De resultaten van de simulatiestudie geven aan dat WAIS-schatters in het algemeen een lage standaardfout hebben en dit onder verschillende omstandigheden. Bovendien is de probabiliteit van de op WAIS-gebaseerde modellen vergelijkbaar met deze van modellen die gebruik maken van alternatieve CoC-schatters. Er worden ook aanbevelingen gegeven over het gebruik van de verschillende CoC-schatters.

Het inpassen van WAIS in de variantiestructuur van een gemengd lineair model levert niet noodzakelijk het best passende model op. Een andere CoC-schatter past soms beter, ondanks het feit dat deze misschien niet PSD is. In dit geval kan een matrixombuigingsroutine gehanteerd worden die toelaat om de matrix om te buigen naar de dichtstbijzijnde PSD matrix. In deze thesis wordt een nieuwe MCMC-gebaseerde ombuigingsmethode voorgesteld en vergeleken met een meer klassieke aanpak die steunt op de singuliere-waardenontbinding van een matrix. Vijf CoC-schatters, waaronder WAIS, de twee bestudeerde matrixombuigingsprocedures en verscheidene andere methodes voor matrixmanipulatie werden geïmplementeerd in het softwarepakket CoCoa, dat gratis ter beschikking wordt gesteld onder de gebruiksvoorwaarden van de GNU Algemene Publieke Licentie.

Initieel worden de geselecteerde fenotypische gegevens geanalyseerd aan de hand van een eenvoudig gemengd lineair model. Dit laat toe om schatters te bekomen voor de omgevingseffecten die gebruikt worden om elke fenotypische meting te corrigeren, zodat een enkelvoudige score voor elke hybride wordt bekomen. Deze scores worden 'willekeurige fenotypes' genoemd en worden vervolgens gebruikt voor het opstellen van een genomisch predictiemodel gebaseerd op ε-SVR en Best Linear Prediction (BLP). Het gebruik van kernelfuncties stelt ε-SVR in staat om een linear model te bepalen in een hoog-dimensionale feature ruimte hetgeen een niet-lineair model vormt in de oorspronkelijke ruimte. De prestaties van de veelzijdige Gaussiaanse kernelfunctie worden onderzocht en er wordt aangetoond dat PSD CoC schatters ook gebruikt kunnen worden als kernelfuncties. De initiële resultaten, die gebruik maken van de RAGT R2n data, geven aan dat ε-SVR en BLP een vergelijkbare accuraatheid van predictie hebben voor verschillende combinaties van merkertypes en agronomische kenmerken. De ε-SVR aanpak laat wel toe om meerdere types predictoren te combineren in hetzelfde model, hetgeen een grotere flexibiliteit oplevert. Desalniettemin is, de via cross-validatie bepaalde accuraatheid van korrelopbrengstpredictie, ontoereikend voor een efficiënte genomische selectie.

In een poging om de accuraatheid te verbeteren, wordt de analyse aan de hand van een

gemengd lineair model herbekeken. Het gebruikte model wordt uitgebreid met additionele vaste omgevingsfactoren en willekeurige G×E-effecten hetgeen betere modellen voor het gemiddelde en de variantie oplevert. Er wordt aangetoond dat een eenvoudige sommatie van BLUP's veel betere trainingsvoorbeelden oplevert in vergelijking met de aanpak die steunt op willekeurige fenotypes. Deze aanpassingen resulteren in cross-validatiegebaseerde accuraatheid die uitzonderlijk hoog is, zelfs wanneer geprobeerd wordt om SCA-waarden te voorspellen. Daarom werd een specifieke veldproef aangelegd waarin de agronomische prestaties van 49 hybriden werden opgemeten in drie locaties in Zuid-Frankrijk. Deze proef laat toe om een definitief oordeel te vellen over de predictiecapaciteit van zowel ε-SVR als BLP.

De resultaten van deze proef tonen aan dat er een aanzienlijke discrepantie is tussen de bepalingen van predictieaccuraatheid door middel van cross-validatie en een effectieve veldproef. De oorzaak van deze discrepantie is echter vrij duidelijk. De trainingsvoorbeelden worden marginaal voorspeld op de effecten van groeiseizoenen en locaties, terwijl de hybriden in de validatieproef conditoneel op een specifiek niveau van deze factoren worden voorspeld. De afwijking tussen beide voorspellingen hangt af van het kenmerk dat onderzocht wordt. Voor korrelopbrengst zijn de correlaties tussen beiden dermate laag dat er weinig hoop is om een betrouwbare genomische selectie uit te voeren. Dit kenmerk heeft dan ook een erg lage heritabiliteit in geavanceerde veredelingsprogramma's wat voornamelijk wordt veroorzaakt door de hoge bijdrage van G×E-effecten tot de totale fenotypische variantie. Deze hoge G×E-variantie zorgt voor de lage correlatie tussen de marginale en conditionele predicties. De twee andere onderzochte kenmerken, namelijk het vochtgehalte van de korrel en het aantal dagen alvorens de bloei, hebben een hogere heritabiliteit en kunnen daardoor beter worden voorspeld door ε-SVR en BLP. De uitersten van de predictieve capaciteiten van deze beide methodes worden verder geanalyseerd door het reduceren van het aantal trainingshybriden en de omvang van de moleculaire vingerafdruk. De accuraatheid van BLP blijkt minder gevoelig aan een reductie van het aantal trainingsvoorbeelden in vergelijking met ε-SVR. Deze laatste is wel beter in het voorspellen van de agronomische resultaten van hybriden wanneer een kleinere of minder informatieve moleculaire vingerafdruk wordt gebruikt.

Curriculum vitae

Name:	Steven Maenhout
Date and place of birth:	August 8, 1977, Eeklo (Belgium)
Address (private):	Gravenstraat 13, B-9970 Kaprijke
Address (work):	Voskenslaan 270, B-9000 Gent
E-mail:	Steven.Maenhout@hogent.be
Telephone number:	0032 (0)9 248 88 60
Fax number:	0032 (0)9 242 42 79

Education

1999	Master in Applied Bio-engineering, option Plant Production, University College Ghent
2003	Predoctoral Training (60 credits), Ghent University
2007	Doctoral Training (60 credits), Ghent University

Publications

International journal publications with peer-review

Maenhout S., De Baets B., Haesaert G. and Van Bockstaele E. (2007). Support vector machine regression for the prediction of maize hybrid performance. *Theoretical and Applied Genetics*, 115:1003-1013.

Maenhout S., De Baets B., Haesaert G. and Van Bockstaele E. (2008). Marker-based screening of maize inbred lines using support vector machine regression. *Euphytica*, 161:123-131.

Maenhout S., De Baets B. and Haesaert G. (2009). Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes. *Theoretical and Applied Genetics*, 118:1181-1192.

Maenhout S., De Baets B. and Haesaert G. (2009). CoCoa: a software tool for estimating the coefficient of coancestry from multilocus genotype data. *Bioinformatics*, 25:2753-2754.

Maenhout S., Haesaert G. and De Baets B. (2010). Prediction of maize single-cross hybrid performance: support vector machine regression versus best linear prediction *Theoretical and Applied Genetics*, 120:415-427.

Presentations at international conferences

19.01.2006	Connectedness of genetic evaluation data, ORBEL 20, 'The Twentieth Conference on Quantitative Methods for Decision Making', Ghent University
31.08.2006	Support Vector Machine Regression for hybrid prediction, XIII Meeting of the Section of Biometrics in Plant Breeding, Zagreb, Croatia
24.08.2007	Support Vector Machine Regression for hybrid maize prediction, The 3rd International Conference on Quantitative Genetics, Zhejiang University, Hangzhou, China
26.09.2008	Kernel based methods for hybrid prediction, Workshop 'Approaches for association mapping and genome-wide genotyping by means of chip hybridization', University of Hohenheim, Stuttgart, Germany
03.09.2009	Marker-based estimation of the coefficient of coancestry in hybrid breeding programmes, XIV Meeting of the Section of Biometrics in Plant Breeding, Dundee, Scotland
09.09.2009	Hybrid prediction through machine learning on commercial maize breeding data, International conference on 'Heterosis in Plants', Genetic and molecular causes and optimal exploitation in breeding, University of Hohenheim, Stuttgart, Germany

Bibliography

A. Gilmour, B. C. and Verbyla, A. (1997). Accounting for natural and extraneous variation in the analysis of field experiments. *Journal of Agricultural, Biological and Environmental Statistics*, 2:269–293.

Anderson, A. D. and Weir, B. S. (2007). A maximum-likelihood method for the estimation of pairwise relatedness in structured populations. *Genetics*, 176:421–440.

Anderson, E. and Brown, W. L. (1952). Origin of corn belt maize and its genetic significance. In Gowen, J. W., editor, *Heterosis*, pages 124–128. Iowa State College Press, Ames.

Aronszajn, N. (1950). Theory of reproducing kernels. *Transactions of the American Mathematical Society*, 68:337–404.

Asahiro, Y., Iwama, K., Tamaki, H., and Tokuyama, T. (2000). Greedily finding a dense subgraph. *Algorithmica*, 34:203–221.

Bahlmann, C., Haasdonk, B., and Burkhardt, H. (2002). On-line handwriting recognition with support vector machines - a kernel approach. In *Proceedings of the 8th International Workshop on Frontiers in Handwriting Recognition*, volume Communication No. 28-12. Vol. 33, pages 49–54. IEEE Computer Society Washington.

Balding, D. J. (2003). Likelihood-based inference for genetic correlation coefficients. *Theoretical Population Biology*, 63:221–230.

Battiti, R. and Protasi, M. (2001). Reactive local search for the maximum clique problem. *Algorithmica*, 29:610–637.

Beal, W. J. (1878). The improvement of grains, fruits, and vegetables. *Michigan State Board of Agriculture Report*, 17:343–345.

Berg, C., Christensen, J. P. R., and Ressel, P. (1984). *Harmonic Analysis on Semigroups: Theory of Positive Definite and Related Functions*. Springer, New York.

Bernardo, R. (1992). Relationship between single-cross performance and molecular heterozygosity. *Theoretical and Applied Genetics*, 83:628–634.

Bernardo, R. (1993). Estimation of coefficient of coancestry using molecular markers in maize. *Theoretical and Applied Genetics*, 85:1055–1062.

Bernardo, R. (1994). Prediction of maize single-cross performance using rflps and information from related hybrids. *Crop Science*, 34:20–25.

Bernardo, R. (1995). Genetic models for predicting maize single-cross performance in unbalanced yield trial data. *Crop Science*, 35:141–147.

Bernardo, R. (1996a). best linear unbiased prediction of the performance of crosses between untested maize inbreds. *Crop Science*, 36:50–56.

Bernardo, R. (1996b). Best linear unbiased prediction of maize single-cross performance given erroneous inbred relationships. *Crop Science*, 36:862–866.

Bernardo, R. (1996c). Best linear unbiased prediction of the performance of crosses between untested maize inbreds. *Crop Science*, 36:872–876.

Bernardo, R., Romero-Severson, J., Ziegle, J., Hauser, J., Joe, L., Hookstra, G., and Doerge, R. W. (2000). Parental contribution and coefficient of coancestry among maize inbreds: pedigree, RFLP and SSR data. *Theoretical and Applied Genetics*, 100:552–556.

Bernardo, R. and Yu, J. (2007). Prospects for genomewide selection for quantitative traits in maize. *Crop Science*, 47:1082–1090.

Bernardo, R. (2008). Molecular markers and selection for complex traits in plants: learning from the last 20 years. *Crop Science*, 48:1649–1664.

Bomze, M., Budinich, M., Pardalos, P. M., and Pelillo, M. (1999). The maximum clique problem. In Du, D. Z. and Pardalos, P. M., editors, *Handbook of Combinatorial Optimization (Supplement Volume A)*, pages 1–74. Kluwer Academic, Dordrecht.

Boser, B. E., Guyon, I. M., and Vapnik, V. (1992). A training algorithm for optimal margin classifiers. In *Proceedings of the 5th Annual ACM Workshop on Computational Learning Theory*, pages 144–152. ACM Press, Pittsburgh.

Brélaz, D. (1979). New methods to color the vertices of a graph. *Communications of the ACM*, 22(4):251–256.

Bueno, J. S. D. and Gilmour, S. G. (2003). Planning incomplete block experiments when treatments are genetically related. *Biometrics*, 59:375–381.

Carraghan, R. and Pardalos, P. M. (1990). An exact algorithm for the maximum clique problem. *Operations Research Letters*, 9:375–382.

Casa, A. M., Pressoir, G., Brown, P. J., Mitchell, S. E., Rooney, W. L., Tuinstra, M. R., Franks, C. D., and Kresovich, S. (2008). Community resources and strategies for association mapping in sorghum. *Crop Science*, 48:30–40.

Castiglioni, P., Ajmone-Marsan, P., van Wijk, R., and Motto, M. (1999). AFLP markers in a molecular linkage map of maize: codominant scoring and linkage group distribution. *Theoretical and Applied Genetics*, 99:425–431.

Chakrabarti, M. C. (1964). On the C-matrix in design of experiments. *Journal of the Indian Statistical Association*, 1:8–23.

Chang, C. C. and Lin, C. J. (2006). *LIBSVM: a library for support vector machines*. `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.

Charcosset, A., Lefort-Buson, M., and Gallais, A. (1991). Relationship between heterosis and heterozygosity at marker loci: a theoretical computation. *Theoretical and Applied Genetics*, 81:571–575.

Charcosset, A. and Essioux, L. (1994). The effect of population-structure on the relationship between heterosis and heterozygosity at marker loci. *Theoretical and Applied Genetics*, 89:336–343.

Charcosset, A., Bonnisseau, B., Touchebeuf, O., Burstin, J., Dubreuil, P., Barriére, Y., Gallais, A., and Denis, J. B. (1998). Prediction of maize hybrid silage performance using marker data: comparison of several models for specific combining ability. *Crop Science*, 38:38–44.

Clauset, A., Shalizi, C. R., and Newman, M. E. J. (2009). Power-law distributions in empirical data. *SIAM Review*, 51:661–703.

Collins, G. N. (1921). Dominance and the vigor of first generation hybrids. *American Naturalist*, 55:116–133.

Comstock, R. E. and Robinson, H. F. (1952). Estimation of the average dominance of genes. In Gowen, J. W., editor, *Heterosis*, pages 494–516. Iowa State College Press, Ames.

Cox, T. S., Kiang, Y. T., Gorman, M. B., and Rodgers, D. M. (1985). Relationship between coefficient of parentage and genetic similarity indices in soybean. *Crop Science*, 25:529–532.

Cristianini, N. and Shawe-Taylor, J. (2000). *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge University Press, Cambridge.

Crow, J. F. (2000). The rise and fall of overdominance. *Plant Breeding Reviews*, 17:225–257.

Cullis, B., Gogel, B., Verbyla, A., and Thompson, R. (1998). Spatial analysis of multi-environment early generation trials. *Biometrics*, 54:1–18.

Darrah, L. L. and Zuber, M. S. (1986). 1985 United States farm maize germplasm base and commercial breeding strategies. *Crop Science*, 26:1109–1113.

Davenport, C. B. (1908). Degeneration, albinism and inbreeding. *Science*, 28:454–455.

De Meyer, H., Naessens, H., and De Baets, B. (2004). Algorithms for computing the min-transitive closure and associated partition tree of a symmetric fuzzy relation. *European Journal of Operational Research*, 155:226–238.

Decoste, D. and Schölkopf, B. (2002). Training invariant support vector machines. *Machine Learning*, 46:161–190.

East, E. M. (1908). Inbreeding in corn. *Reports of the Connecticut Agricultural Experiments Station for 1907*, pages 419–428.

East, E. M. and Jones, D. F. (1919). *Inbreeding and Outbreeding: their Genetic and Sociological Significance*. J.B. Lippincott Company, Philadelphia.

East, E. M. (1936). Heterosis. *Genetics*, 21:375–397.

Emik, L. and Terrill, C. (1949). Systematic procedures for calculating inbreeding coefficients. *Journal Heredity*, 40:51–55.

Erkut, E. (1990). The discrete p-dispersion problem. *European Journal of Operational Research*, 46:48–60.

Erkut, E., Ulkusal, Y., and Yenicerioglu, O. (1994). A comparison of p-dispersion heuristics. *Computers & Operations Research*, 21(10):1103–1113.

Evgeniou, T., Pontil, M., and Poggio, T. (1999). A unified framework for regularization networks and support vector machines. Technical Report AIM-1654, Massachusetts Institute of Technology. `http://hdl.handle.net/1721.1/7261`.

Evgeniou, T., Poggio, T., Pontil, M., and Verri, A. (2002). Regularization and statistical learning theory for data analysis. *Computational Statistics & Data Analysis*, 38:421–432.

Fan, R. E., Chen, P. H., and Lin, C. J. (2005). Working set selection using second order information for training support vector machines. *Journal of Machine Learning*, 6:1889–1918.

Feige, U. and Seltser, M. (1997). On the densest k-subgraph problem. Technical Report CS97-16, The Weizmann Institute, Rehovot, Israel. `http://citeseer.ist.psu.edu/feige97densest.html`.

Foulley, J. L., Bouix, J., Goffinet, B., and Elsen, J. M. (1990). Connectedness in genetic evaluation. In Gianola, D. and Hammond, K., editors, *Advances in Statistical Methods for Genetic Improvement of Livestock*, pages 277–308. Springer-Verlag, Heidelberg.

Frisch, M., Thiemann, A., Fu, J., Schrag, T. A., Scholten, S., and Melchinger, A. E. (2010). Transcriptome-based distance measures for grouping of germplasm and prediction of hybrid performance in maize. *Theoretical and Applied Genetics*. doi=10.1007/s00122-009-1204-1.

Galassi, M., Davies, J., Theiler, J., Gough, B., Priedhorsky, R., Jungman, G., Booth, M., and F.Rossi (2009). *GNU Scientific Library Reference Manual - Third Edition (v1.12)*. `http://www.gnu.org/software/gsl/`.

Gianola, D., Fernando, R. L., and Stella, A. (2006). Genomic-assisted prediction of genetic value with semiparametric procedures. *Genetics*, 173:1761–1776.

Gianola, D. and van Kaam, J. B. C. H. M. (2008). Reproducing kernel Hilbert spaces regression methods for genomic assisted prediction of quantitative traits. *Genetics*, 178:2289–2303.

Gilmour, A., Gogel, B., Cullis, B., Welham, S., and Thompson, R. (2002). *ASREML User Guide Release 1.0*. VSN International Ltd.

Godshalk, E. B., Lee, M., and Lamkey, K. R. (1990). Relationship of restriction fragment length polymorphisms to singlecross hybrid performance of maize. *Theoretical and Applied Genetics*, 80:273–280.

González-Recio, O., Gianola, D., Long, N., Weigel, K. A., Rosa, G. J. M., and Avendaño, S. (2008). Nonparametric methods for incorporating genomic information into genetic evaluations: an application to mortality in broilers. *Genetics*, 178:2305–2313.

Gower, J. C. (1971). A general coefficient of similarity and some of its properties. *Biometrics*, 27:857–874.

Gower, J. C. and Legendre, P. (1986). Metric and Euclidean properties of dissimilarity coefficients. *Journal of Classification*, 3:5–48.

Guyon, I., Weston, J., Barnhil, S., and Vapnik, V. (2002). Gene selection for cancer classification using support vector machines. *Machine Learning*, 46:389–422.

Haasdonk, B. and Keysers, D. (2002). Tangent distance kernels for support vector machines. In *Proceedings of the 16th International Conference on Pattern Recognition*.

Hayes, B. J., Bowman, P. J., Chamberlain, A. J., and Goddard, M. E. (2009). Genomic selection in dairy cattle: progress and challenges. *Journal of Dairy Science*, 92:433–443.

Hayes, J. F. and Hill, W. G. (1981). Modification of estimates of parameters in the construction of genetic selection indices ('Bending'). *Biometrics*, 37:483–493.

Heiligers, B. (1991). A note on connectedness of block designs. *Metrika*, 38:377–381.

Henderson, C. R. (1975). Best linear unbiased estimation and prediction under a selection model. *Biometrics*, 31:423–447.

Henderson, C. R. (1984). *Applications of Linear Models in Animal Breeding*. University of Guelph Press, Guelph.

Henshall, J. M. and Meyer, K. (2002). Pdmatrix–programs to make matrices positive definite. In *Proceedings of the 7th World Congress on Genetics Applied to Livestock Production*, volume Communication No. 28-12. Vol. 33, pages 753–754.

Hepler, A. B. (2005). *Improving forensic identification using Bayesian networks and relatedness estimation*. PhD thesis, North-Carolina State University, Raleigh, NC.

Ho, J. C., Kresovich, S., and Lamkey, K. R. (2005). Extent and distribution of genetic variation in U.S. maize: historically important lines and their open-pollinated dent and flint progenitors. *Crop Science*, 45:1891–1900.

Hochholdinger, F. and Hoecker, N. (2007). Towards the molecular basis of heterosis. *Trends in Plant Science*, 12:427–432.

Hsu, C. W., Chang, C. C., and Lin, C. J. (2003). A practical guide to support vector classification. `http://www.csie.ntu.edu.tw/~cjlin/papers/guide/guide.pdf`.

Hull, F. H. (1945). Recurrent selection for specific combining ability in corn. *Journal of the American Society of Agronomy*, 37:134–145.

Jacquard, A. (1974). *The Genetic Structure of Populations*. Springer-Verlag, New York.

Jannink, J. L., Bink, M. C. A. M., and Jansen, R. C. (2001). Using complex plant pedigrees to map valuable genes. *Trends in Plant Science*, 6:337–342.

Jensen, T. and Toft, B. (1995). *Graph Coloring Problems*. Wiley, New York.

Jones, D. F. (1917). Dominance of linked factors as a means of accounting for heterosis. *Genetics*, 2:466–479.

Jones, D. R., Schonlau, M., and Welch, W. J. (1998). Efficient global optimization of expensive black-box functions. *Journal of Global Optimization*, 13:455–492.

Jorjani, H., Klei, L., and Emanuelson, U. (2003). A simple method for weighted bending of genetic (co)variance matrices. *Journal of Dairy Science*, 86:677–679.

Keerthi, S. S., Shevade, S. K., Bhattacharyya, C., and Murty, K. R. K. (2001). Improvements to Platt's SMO algorithm for SVM classifier design. *Neural Computation*, 13:637–649.

Kennedy, B. W. and Trus, D. (1993). Considerations on genetic connectedness between management units under an animal model. *Journal of Animal Science*, 71:2341–2352.

Khanna, K. R. (1991). *Biochemical Aspects of Crop Improvement*. CRC-Press, Florida.

Labate, J. A., Lamkey, K. R., Mitchell, S. E., Kresovich, S., Sullivan, H., and Smith, S. C. (2003). Molecular and historical aspects of corn belt dent diversity. *Crop Science*, 43:80–91.

Laloé, D. (1993). Precision and information in linear models of genetic evaluation. *Genetics Selection Evolution*, 25:557–576.

Laloé, D., Phocas, F., and Ménissier, F. (1996). Considerations about measures of precision and connection in mixed linear models of genetic evaluation. *Genetics Selection Evolution*, 28:359–378.

Lee, E. A. and Tracey, W. F. (2009). Modern maize breeding. In Bennetzen, J. L. and Hake, S., editors, *Handbook of Maize: Genetics and Genomics*, pages 141–160. Springer.

Lescovec, J. and Faloutsos, C. (2006). Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 631–636. ACM, New York.

Li, C. C., Weeks, D. E., and Chakravarti, A. (1993). Similarity of DNA fingerprints due to chance and relatedness. *Human Heredity*, 43:45–52.

Loiselle, B. A. and Graham, V. L. S. J. N. C. (1995). Spatial genetic structure of a tropical understory shrub, *psychotria officinalis* (rubiaceae). *American Journal of Botany*, 82:1420–1425.

Lu, H. and Bernardo, R. (2001). Molecular marker diversity among current and historical maize inbreds. *Theoretical and Applied Genetics*, 103:613–617.

Lynch, M. (1988). Estimation of relatedness by DNA fingerprinting. *Molecular Biology and Evolution*, 5:584–599.

Lynch, M. and Walsh, B. (1998). *Genetics and Analysis of Quantitative Traits*. Sinauer Associates, Inc, Sunderland.

Lynch, M. and Ritland, K. (1999). Estimation of pairwise relatedness with molecular markers. *Genetics*, 152:1753–1766.

Melchinger, A. E., Lee, M., Lamkey, K. R., and Woodman, W. L. (1990). Genetic diversity for restriction fragment length polymorphisms: relation to estimated genetic effects in maize inbreds. *Crop Science*, 30:1033–1040.

Melchinger, A. E. (1999). Genetic diversity and heterosis. In Coors, J. G. and Pandey, S., editors, *The Genetics and Exploitation of Heterosis in Crops*, pages 99–118. American Society of Agronomy, Madison.

Mercer, J. (1909). Functions of positive and negative type and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, 209:415–446.

Meuwissen, T. H. E., Hayes, B. J., and Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, 157:1819–1829.

Meza, J. C., Oliva, R. A., Hough, P. D., and Williams, P. J. (2007). OPT++: an object-oriented toolkit for nonlinear optimization. *ACM transactions on Mathematical Software*, 33(2):12:1–12:27.

Milligan, B. G. (2002). Maximum-likelihood estimation of relatedness. *Genetics*, 163:1153–1167.

Naessens, H., De Meyer, H., and De Baets, B. (2002). Algorithms for the computation of T-transitive closures. *IEEE Transactions on Fuzzy Systems*, 10:541–551.

Nelder, J. A. and Mead, R. (1965). A simplex method for function minimization. *Computer Journal*, 7(4):308–313.

Nocedal, J. and Wright, S. J. (1999). *Numerical Optimization*. Springer-Verlag, New York.

Oakey, H., Verbyla, A. P., Cullis, B. R., Wei, X., and Pitchford, W. S. (2007). Joint modeling of additive and non-additive (genetic line) effects in multi-environment trials. *Theoretical and Applied Genetics*, 114:1319–1332.

Östergård, P. R. J. (2002). A fast algorithm for the maximum clique problem. *Discrete Applied Mathematics*, 120:197–207.

Osuna, E., Freund, R., and Girosi, F. (1997). Training support vector machines: an application to face detection. In *Proceedings of the 1997 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 130–136. IEEE Computer Society, Washington.

Panter, D. M. and Allen, F. L. (1995a). Using best linear unbiased predictions to enhance breeding for yield in soybean. 1. Choosing parents. *Crop Science*, 35:397–405.

Panter, D. M. and Allen, F. L. (1995b). Using best linear unbiased predictions to enhance breeding for yield in soybean. 2. Selection of superior crosses from a limited number of yield trials. *Crop Science*, 35:405–410.

Patterson, H. D. and Thompson, R. (1971). Recovery of inter-block information when block sizes are equal. *Biometrika*, 58:545–554.

Piepho, H. P., Möhring, J., Melchinger, A. E., and Büchse, A. (2008). BLUP for phenotypic selection in plant breeding and variety testing. *Euphytica*, 161:209–228.

Platt, J. C. (1998). Sequential minimal optimization: a fast algorithm for training support vector machines. Technical Report MSR-TR-98-14, Microsoft Research. `http://citeseer.ist.psu.edu/platt98sequential.html`.

Pritchard, J. K., Stephens, M., and Donnelly, P. (2000). Inference of population structure using multilocus genotype data. *Genetics*, 155:945–959.

Queller, D. C., , and Goodnight, K. F. (1989). Estimating relatedness using genetic markers. *Evolution*, 43:258–275.

Rao, C. R. (1952). *Advanced Statistical Methods in Biometric Research*. John Wiley & Sons Inc., New York.

Ravi, S., Rosenkrantz, D., and Tayi, G. (1991). Facility dispersion problems: heuristics and special cases. *Lecture Notes in Computer Science*, 519:355–66.

Remington, D. L., Thornsberry, J. M., Matsuoka, Y., Wilson, L. M., Whitt, S. R., Doebley, J., S. Kresovich, M. M. G., and Buckler, E. S. (2001). Structure of linkage disequilibrium and phenotypic associations in the maize genome. *Proceedings of the National Academy of Sciences of the United States of America*, 98:11479–11484.

Ritland, K. (1996). Estimators for pairwise relatedness and individual inbreeding coefficients. *Genetical Research*, 67:175–185.

Rogers, J. S. (1972). Measures of genetic similarity and genetic distance. In *Studies in genetics volume 7*, pages 145–153. University of Texas Publications 7213.

Schölkopf, B., Smola, A. J., Williamson, R. C., and Bartlett, P. L. (2000). New support vector algorithms. *Neural Computation*, 12:1207–1245.

Schölkopf, B. and Smola, A. J. (2001). *Learning with Kernels, Support Vector Machines, Regularization, Optimization, and Beyond*. The MIT Press.

Schrag, T. A., Melchinger, A. E., Sørensen, A. P., and Frisch, M. (2006). Prediction of single-cross hybrid performance for grain yield and grain dry matter content in maize

using AFLP markers associated with QTL. *Theoretical and Applied Genetics*, 113:1037–1047.

Schrag, T. A., Maurer, H. P., Melchinger, A. E., Piepho, H. P., Peleman, J., and Frisch, M. (2007). Prediction of single-cross hybrid performance in maize using haplotype blocks associated with QTL for grain yield. *Theoretical and Applied Genetics*, 114:1345–1355.

Schrag, T. A., Möhring, J., Maurer, H. P., Dhillon, B. S., Melchinger, A. E., Piepho, H. P., Sørensen, A. P., and Frisch, M. (2009). Molecular marker-based prediction of hybrid performance in maize using unbalanced data from multiple experiments with factorial crosses. *Theoretical and Applied Genetics*, 118:741–751.

Schrag, T. A., Möhring, J., Kusterer, B., Dhillon, B. S., Melchinger, A. E., Piepho, H. P., and Frisch, M. (2010). Prediction of hybrid performance in maize using molecular markers and joint analyses of hybrids and parental inbreds. *Theoretical and Applied Genetics*. doi=10.1007/s00122-009-1208-x.

Shawe-Taylor, J. and Cristianini, N. (2004). *Kernel Methods for Pattern Analysis*. Cambridge University Press, Cambridge.

Shull, G. H. (1908). The composition of a field of maize. *Annual Report of the American Breeders Association*, 4:296–301.

Shull, G. H. (1911). The genotypes of maize. *American Naturalist*, 45:234–252.

Shull, G. H. (1914). Duplicate genes for capsule-form inbursa bursa-pastoris. *Zeitschrift für Induktive Abstammungs- und Vererbungslehre*, 12:297–149.

Shull, G. H. (1948). What is "heterosis"? *Genetics*, 33:439–446.

Smith, A., Cullis, B., and Thompson, R. (2001). Analyzing variety by environment data using multiplicative mixed models and adjustments for spatial field trend. *Biometrics*, 57:1138–1147.

Smith, J. S. C. and Smith, O. S. (1989). The use of morphological, biochemical, and genetic characteristics to measure distance and to test for minimum distance between inbred lines of maize (zea mays l.). *Mimeo of paper presented at UPOV Workshop, Versailles, France, October 1989. Pioneer Hi-Bred International, Inc., Johnston, IA.*

Sørensen, A. C., Pong-Wong, R., Windig, J. J., and Woolliams, J. A. (2002). Precision of methods for calculating identity-by-descent matrices using multiple markers. 34:557–579.

Sprague, G. F. and Tatum, L. A. (1948). General vs. specific combining ability in single crosses of corn. *Journal of the American Society of Agronomy*, 34:923–932.

Sprague, G. F. and Russell, W. A. (1956). Some evidence on type of gene action involved in yield heterosis in maize. In *Proceedings of the International Genetics Symposia, Tokyo & Kyoto*, pages 522–526.

Sprague, G. F. (1983). Heterosis in maize: Theory and practice. In Frankel, R., editor, *Heterosis: Reappraisal of Theory and Practice (Monographs on Theoretical and Applied Genetics)*, pages 47–70. Springer-Verlag, Berlin.

Stich, B., Melchinger, A. E., Frisch, M., Maurer, H. P., Heckenberger, M., and Reif, J. C. (2005). Linkage disequilibrium in european elite maize germplasm investigated with SSRs. *Theoretical and Applied Genetics*, 111:723–730.

Stich, B., Maurer, H. P., Melchinger, A. E., Frisch, M., Heckenberger, M., Rouppe van der Voort, J., Peleman, J., Sørensen, A. P., and Reif, J. C. (2006). Comparison of linkage disequilibrium in elite European maize inbred lines using AFLP and SSR markers. *Molecular Breeding*, 17:217–226.

Stich, B., Melchinger, A. E., Piepho, H. P., Hamrit, S., Schipprack, W., Maurer, H. P., and Reif, J. C. (2007). Potential causes of linkage disequilibrium in a european maize breeding program investigated with computer simulations. *Theoretical and Applied Genetics*, 115:529–536.

Stich, B., Möhring, J., Piepho, H. P., Heckenberger, M., Buckler, E. S., and Melchinger, A. E. (2008). Comparison of mixed-model approaches for association mapping. *Genetics*, 178:1745–1754.

Stuber, C. W. and Cockerham, C. C. (1966). Gene effects and variances in hybrid populations. *Genetics*, 54:1279–1286.

Suykens, J. A. K. and Vandewalle, J. (1999). Least squares support vector machine classifiers. *Neural Processing Letters*, 3:293–300.

Thompson, E. A. (1975). The estimation of pairwise relationships. *Annals of Human Genetics*, 39:173–188.

Tikhonov, A. N. and Arsenin, V. Y. (1977). *Solution of Ill-Posed Problems*. Winston & Sons, Washington DC.

Tomita, E. and Seki, T. (2003). An efficient branch-and-bound algorithm for finding a maximum clique. *Discrete Mathematics and Theoretical Computer Science*, 2731:278–289.

Tracey, W. F. and Chandler, M. A. (1988). The historical and biological basis of the concept of heterotic patterns in corn belt dent maize. In Lamkey, K. R. and Lee, M., editors, *Plant breeding: the Arnel R. Hallauer International Symposium*, pages 219–233. Blackwell Publishing.

Van de Casteele, T., Galbusera, P., and Matthysen, E. (2001). A comparison of microsatellite-based pairwise relatedness estimators. *Molecular Ecology*, 10:1539–1549.

Vapnik, V. N. and Chervonenkis, A. Y. (1971). On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 17:264–280.

Vapnik, V. N. (1995). *The Nature of Statistical Learning Theory*. Springer-Verlag, New-York.

Vapnik, V. N. (1998). *Statistical Learning Theory*. Wiley, New York.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., van de Lee, T., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., and Zabeau, M. (1995). AFLP: a new technique for DNA-fingerprinting. *Nucleic Acids Research*, 23:4407–4414.

Vuylsteke, M., Mank, R., Antonise, R., Bastiaans, E., Senior, M. L., Stuber, C. W., Melchinger, A. E., Lübberstedt, T., Xia, X. C., Stam, P., Zabeau, M., and Kuiper, M. (1999). Two high-density AFLP (R) linkage maps of zea mays l.: analysis of distribution of AFLP markers. *Theoretical and Applied Genetics*, 99:921–935.

Vuylsteke, M., Kuiper, M., and Stam, P. (2000). Chromosomal regions involved in hybrid performance and heterosis: their AFLP R -based identification and practical use in prediction models. *Heredity*, 85:208–218.

Wahba, G. (1990). *Spline Models for Observational Data (CBMS-NSF Regional Conference Series in Applied Mathematics)*. SIAM: Society for Industrial and Applied Mathematics.

Wang, J. (2002). An estimator for pairwise relatedness using molecular markers. *Genetics*, 160:1203–1215.

Warshall, S. (1962). A theorem on boolean matrices. *Journal of the ACM*, 9:11–12.

Welham, S., Cullis, B., Gogel, B., Gilmour, A., and Thompson, R. (2004). Prediction in linear mixed models. *Australian & New Zealand Journal of Statistics*, pages 325–347.

Weston, J., Mukherjee, S., Chapelle, O., Pontil, M., Poggio, T., and Vapnik, V. (2000). Feature selection for SVMs. *Advances in Neural Information Processing Systems*, 13:668–674.

Wright, S. (1943). Isolation by distance. *Genetics*, 28:114–138.

Wright, S. (1951). The genetical structure of populations. *Annals of Eugenics*, 15:323–354.

Yan, J. B., Shah, T., Warburton, M. L., Buckler, E. S., McMullen, M. D., and Crouch, J. (2009). Genetic characterization and linkage disequilibrium estimation of a global maize collection using SNP markers. *PLOS ONE*, 4:e8451.

Yu, J., Pressoir, G., Briggs, W. H., Vroh Bi, I., Yamasaki, M., Doebley, J. F., McMullen, M. D., Gaut, B. S., Nielsen, D. M., Holland, J. B., Kresovich, S., and Buckler, E. S. (2006). A unified mixed-model method for association mapping that accounts for multiple levels of relatedness. *Nature Genetics*, 38:203–208.

Zhang, Z., Todhunter, R. J., Buckler, E. S., and Van Vleck, L. D. (2007). Technical note: use of marker-based relationships with multiple-trait derivative-free restricted maximal likelihood. *Journal of Animal Science*, 85:881–885.

Zien, A., Rätsch, G., Mika, S., Schölkopf, B., Lengauer, T., and Muller, K. R. (2000). Engineering support vector machine kernels that recognize translation initiation sites. *Bioinformatics*, 16:799–807.