Theoretical Population Biology 132 (2020) 16–23

Contents lists available at ScienceDirect

Theoretical Population Biology

journal homepage: www.elsevier.com/locate/tpb

On the approximation of interaction effect models by Hadamard powers of the additive genomic relationship

Johannes W.R. Martini *, Fernando H. Toledo, José Crossa

International Maize and Wheat Improvement Center (CIMMYT), Mexico

ARTICLE INFO

Article history:
Received 14 September 2019
Available online 25 January 2020

Keywords:
Epistasis
Genetic interaction
Hadamard power

ABSTRACT

Whole genome epistasis models with interactions between different loci can be approximated by genomic relationship models based on Hadamard powers of the additive genomic relationship. We illustrate that the quality of this approximation reduces when the degree of interaction d increases. Moreover, considering relationship models defined as weighted sum of interactions of different degree, we investigate the impact of this decreasing quality of approximation of the summands on the approximation of the weighted sum. Our results indicate that these approximations remain on a reliable level, but their quality reduces when the weights of interactions of higher degrees do not decrease quickly.

1. Introduction

With the broad availability of genomic data of individual animals or plant lines, genomic prediction (Meuwissen et al., 2001) has been widely implemented in modern breeding programs (Hayes et al., 2009; Jannink et al., 2010; Meuwissen et al., 2016; Crossa et al., 2017). The standard method *genomic best linear unbiased prediction* (GBLUP) is based on the commonly used additive effect model (Falconer and Mackay, 1996; Gianola et al., 2009). Given the $n \times p$ matrix $\mathbf{M}$ describing the marker states of the n individuals at p loci, the additive effect model is defined by

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{M}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{1}$$

Here, $\mathbf{y}$ is the $n \times 1$ vector of phenotypic data, $\mathbf{1}_n$ an $n \times 1$ vector with each entry equal to 1, μ a fixed effect, $\boldsymbol{\beta}$ a $p \times 1$ vector of marker effects and $\boldsymbol{\epsilon}$ an $n \times 1$ vector of errors. Moreover, usually the additional assumptions of $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{I}_p)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_n)$ are made. $\mathbf{I}_p$ and $\mathbf{I}_n$ denote the Identity matrix of respective dimension (Note that the term GBLUP usually refers to a reformulated version of Eq. (1) using $\mathbf{g} := \mathbf{M}\boldsymbol{\beta}$, but this distinction will not be used here since both models are statistically equivalent).

Having estimated/predicted the relevant parameters as $\hat{\mu}$ and $\hat{\boldsymbol{\beta}}$, the predicted effect of a change at an arbitrary locus k is independent of the state of other markers. This characteristic seems contrary to the function of biological systems which rely on interaction and where thus the effect of a change at locus k is assumed to depend on the genetic background. The discrepancy between the intrinsic logic of the statistical additive model and the mechanistics of biological processes may provide a motivation to consider "non-additive" relationships for the prediction of (non-additive) total genetic values or phenotypes (de los Campos et al., 2009; Ober et al., 2011). An epistasis model extending the additive setup of Eq. (1) with products of markers as additional predictors (Ober et al., 2015; Jiang and Reif, 2015; Martini et al., 2016) is called the *extended genomic best linear unbiased prediction* (EGBLUP) (Jiang and Reif, 2015). In more detail, this pair epistasis model is defined by

$$y_i = \mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} + \sum_{k=1,\dots,p-1;l>k} M_{i,k}M_{i,l}h_{k,l} + \epsilon_i. \tag{2}$$

Here, μ and $\boldsymbol{\beta}$ are as previously defined and $\mathbf{M}_{i,\bullet}$ denotes the ith row of $\mathbf{M}$, that is the genomic data of individual i. Moreover, $h_{k,l}$ is the interaction effect of loci k and l with $h_{k,l} \overset{i.i.d}{\sim} \mathcal{N}(0, \sigma^2_h)$ and all random effects being stochastic independent from each other. It has been demonstrated that model (2) is equivalent (Martini et al., 2016) to a model

$$\mathbf{y} = \mathbf{1}_n\mu + \mathbf{g}_1 + \mathbf{g}_2 + \boldsymbol{\epsilon} \tag{3}$$

with $\mathbf{g}_1 \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{G})$ and $\mathbf{g}_2 \sim \mathcal{N}(\mathbf{0}, \sigma^2_h \mathbf{H}^{(2)})$ and where $\mathbf{G} := \mathbf{M}\mathbf{M}'$ and

$$\mathbf{H}^{(2)} := 0.5(\mathbf{G} \circ \mathbf{G}) - 0.5(\mathbf{M} \circ \mathbf{M})(\mathbf{M} \circ \mathbf{M})'. \tag{4}$$

The operator $\circ$ denotes here the Hadamard, that is the entry-wise product. This model and some variations have been shown to be able to increase predictive ability for some incidences and compared to the additive GBLUP model (Su et al., 2012; Ober et al., 2015; Jiang and Reif, 2015; Martini et al., 2016). Moreover, these

* Corresponding author.
E-mail address: j.martini@cgiar.org (J.W.R. Martini).

https://doi.org/10.1016/j.tpb.2020.01.004

types of relationship matrices have also been used to control genetic background effects in association studies (Xu, 2013).

In Eq. (2), the interactions are modeled pairwise and only between different loci ($l > k$). Some variations of this model have been used in literature (Jiang and Reif, 2015; Martini et al., 2016) defined by allowing interaction of the loci with themselves ($l \geq k$)

$$y_i = \mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} \quad + \sum_{k=1,\ldots,p;l\geq k} M_{i,k}M_{i,l}h_{k,l} + \epsilon_i. \tag{5}$$

or modeling the p^2 interactions by counting the interaction between different loci twice ($k, l = 1, \ldots, p$)

$$y_i = \mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} \quad + \sum_{k,l=1,\ldots,p} M_{i,k}M_{i,l}h_{k,l} + \epsilon_i. \tag{6}$$

It has been shown that the interaction terms of Eqs. (5)–(6) translate to covariance matrices of $\mathbf{g}_2$ of Eq. (3) the following way (Martini et al., 2016):

$$\sum_{k=1,\ldots,p;l\geq k} M_{i,k}M_{i,l}h_{k,l} \quad \widehat{=} \quad 0.5(\mathbf{G}\circ\mathbf{G}) + 0.5(\mathbf{M}\circ\mathbf{M})(\mathbf{M}\circ\mathbf{M})' \tag{7}$$

$$\sum_{k,l=1,\ldots,p} M_{i,k}M_{i,l}h_{k,l} \quad \widehat{=} \quad (\mathbf{G}\circ\mathbf{G}) \tag{8}$$

Moreover, it has also been demonstrated that for higher degrees of interactions d, the sum of all p^d d-wise interaction terms translate to Hadamard powers of $\mathbf{G}$ (Martini et al., 2016):

$$\sum_{k,l,m=1,\ldots,p} M_{i,k}M_{i,l}M_{i,m}h_{k,l,m} \quad \widehat{=} \quad (\mathbf{G}\circ\mathbf{G}\circ\mathbf{G}), \tag{9}$$

$$\sum_{k,l,m,o=1,\ldots,p} M_{i,k}M_{i,l}M_{i,m}M_{i,o}h_{k,l,m,o} \quad \widehat{=} \quad (\mathbf{G}\circ\mathbf{G}\circ\mathbf{G}\circ\mathbf{G}), \tag{10}$$

and analogously for any degree d. We are not aware of a general concise formula that generalizes Eq. (4) to a general higher degree d (the cases $d = 3$ and $d = 4$ are treated in the Appendix.) Note here that Eq. (9) also includes terms of form $M_{i,k}^3$ that is a three way interaction within a locus. Terms of this type of intra-locus interactions of higher degree d may be difficult to interpret from a quantitative genetics point of view.

It has been argued in the context of a specific marker coding (VanRaden, 2008) that for increasing number of markers p, the quality of an approximation of Eq. (4), which models only the interactions between different loci, by Eq. (8), which models p^2 interactions, improves (Jiang and Reif, 2015). The reason for this improving quality of approximation is roughly spoken a result of $(\mathbf{M}\circ\mathbf{M})(\mathbf{M}\circ\mathbf{M})'$ getting relatively small compared to $(\mathbf{G}\circ\mathbf{G})$, and the factor 0.5 being compensated by an adapted estimate of the variance component $\hat{\sigma}_h^2$.

For the case of $d = 2$, Eq. (4) allows to use model (2) directly. However, when interested in a model with interactions of higher degree between different loci, an approximation by Hadamard products of the additive relationship matrix would immensely reduce computational effort. Hadamard powers of $\mathbf{G}$ are easily calculated, whereas a matrix $\mathbf{H}^{(d)}$ modeling interactions only between different loci is difficult to calculate for larger d. Given the lack of a general concise formula analogous to Eq. (4) for general degree d, a straight-forward but computationally expensive approach to derive $\mathbf{H}^{(d)}$ is to calculate a matrix of all possible products of d columns and multiply it with its transpose. This task can quickly become very difficult since the binomial coefficient $\binom{p}{d}$ grows very fast.

We show that the argumentation which illustrates that for $d = 2$ and increasing p the quality of an approximation of Eq. (4) by Eq. (8) improves, can analogously be adapted to any fixed degree d and increasing p. This means that for any fixed degree of interaction d and increasing p, the quality of the approximation of a model based on interactions between different loci by a Hadamard power of the additive genomic relationship improves.

A different situation – which is important for limit considerations of models with increasing degree of interaction – is increasing d for fixed p. The incorporation of higher degree interactions can lead to new relationship models that aim at reflecting biological complexity better. We show that the quality of the approximation of a model with interactions between different loci by Hadamard powers of $\mathbf{G}$ reduces when p is fixed and d increases. Moreover, we investigate the limit behavior of weighted sums of interactions of increasing degree, in particular, their reliability when substituting interactions between different loci by Hadamard powers of the additive relationship. We show that the approximation may be less reliable if the weights of higher degree interactions do not decrease fast enough.

As a remark please be aware of the problem that the coding of the markers has an impact on epistasis models which use the products of marker values as additional predictor variables (He and Parida, 2016; Martini et al., 2017, 2019). However, this topic of how to code markers will be ignored in this manuscript. The presented results are independent of the coding of markers.

2. Degree d = 2

2.1. Some words on an improvement of the approximation

Let us first make some theoretical considerations on what "the quality of an approximation of Eq. (4) by Eq. (8) improves" shall mean. We define

$$\tilde{\mathbf{H}}_p^{(2)} := \mathbf{G}_p \circ \mathbf{G}_p \quad \text{and}$$
$$\mathbf{H}_p^{(2)} := 0.5\mathbf{G}_p \circ \mathbf{G}_p - 0.5(\mathbf{M}_p \circ \mathbf{M}_p)(\mathbf{M}_p \circ \mathbf{M}_p)'$$

$\mathbf{G}_p$ denotes here the additive genomic relationship matrix based on p markers. Considering the fact that a factor $c \in \mathbb{R}^+$ can be compensated by estimating a different variance component, a first idea to make the statement on improving the quality of the approximation more precise could be:

$$\exists\, c \in \mathbb{R}^+ \text{ such that } \lim_{p\to\infty} \tilde{\mathbf{H}}_p^{(2)} = c \lim_{p\to\infty} \mathbf{H}_p^{(2)}. \tag{11}$$

This expression has earlier been used (Jiang and Reif, 2015) with $c = 2$ and for the specific situation of an allele-frequency-centered and scaled additive genomic relationship according to VanRaden (2008). (Note again that even though subtracting the mean from each column will not have an effect on the prediction of additive effects (Strandén and Christensen, 2011; Martini et al., 2017), it has been shown that this transformation has an impact if the centered values are multiplied to model interactions (He and Parida, 2016; Martini et al., 2017, 2019).)

In the current setup – with $\mathbf{G} := \mathbf{M}\mathbf{M}'$ – expression (11) raises some questions and would require case distinctions or restrictions. Additionally to the formal question of which metric to use to define the limit, there are more conceptual questions. For instance it is not clear how it should be decided which marker pattern a new column has when another marker column is added. Without a restriction on how to add a new column, one can find examples for which the limits in Eq. (11) are not defined or for which a convergence of the entries to 0 leads to a situation in which Eq. (11) is satisfied, but the approximation of the two matrices does not improve. Some examples can be found in the Appendix.

2.2. A simple criterion for the quality of the approximation

To avoid these complications, we choose a simple way to characterize how good the approximation is for fixed p and $d = 2$. This criterion will not solve the problem of how to add new marker columns when p increases, but gives a simple and well-defined equation.

The model described in Eq. (8) models p^2 interactions consisting of the $\binom{p}{2}$ interactions which we want to model (each of them twice) and additionally the p interactions of markers with themselves which are not included in Eq. (4). Since model (4) is a submodel of $\mathbf{G} \circ \mathbf{G}$, and since each interaction has the same influence on the relationship matrix due to assuming their effects to be independent and identically distributed, we can *define* a measure for the "error" of the approximation as the proportion of interactions which we model in $\mathbf{G} \circ \mathbf{G}$ but which are not included in Eq. (4).

In this case of $d = 2$, this is given by the p^2 interactions which we model minus twice the $\binom{p}{2}$ interactions included in Eq. (4).

$$\mathrm{E}_2(p) := \frac{p^2 - 2!\binom{p}{2}}{p^2} = \frac{p}{p^2} = \frac{1}{p}. \tag{12}$$

E_2 describes the error as portion of interactions which we model in our approximation but which are not included in Eq. (4). We see that

$$\lim_{p\to\infty} \mathrm{E}_2(p) = 0,$$

which confirms relatively easily that $\mathbf{G} \circ \mathbf{G}$ is a good approximation for Eq. (4) if p is large. Note here that the difference between Eqs. (2) and (6) which is here – for degree 2 – (not) including interactions of a marker with itself, can also be considered as the difference of the possible events when drawing from $\{1, \ldots, p\}$ with or without replacement. We have to subtract from a set of events of drawing with replacement, those that are not possible when drawing without replacement. This view may facilitate to comprehend the relation between the two models when considering higher degrees d afterwards.

Expression $\mathrm{E}_2(p)$ is well-defined, but also only guarantees that the approximation of $\mathbf{H}_p^{(2)}$ by $\tilde{\mathbf{H}}_p^{(2)}$ improves, if the marker data behaves sufficiently randomly. The problem of how to add new columns when a limit behavior is investigated remains. Consider for instance the sequence $\mathbf{M}_p$ with the first column $\begin{pmatrix}1\\0\end{pmatrix}$ and all other entries 0. $\tilde{\mathbf{H}}_p^{(2)}$ will remain constantly $\begin{pmatrix}1 & 0\\0 & 0\end{pmatrix}$ whereas $\mathbf{H}_p^{(2)}$ is constantly the $\mathbf{0}_{2\times 2}$ matrix. Thus, the approximation will not improve. The reason is here that the only interaction which is not zero is the interaction of the first marker with itself for individual 1. An improving approximation will only be given if the values of $M_{i,k}^2$ do not behave too differently from the values of $M_{i,k}M_{i,l}$.

3. General degree *d*

Analogously, for general $d \leq p$, Eq. (12) generalizes to

$$\mathrm{E}_d(p) := \frac{p^d - d!\binom{p}{d}}{p^d} = 1 - \frac{d!\binom{p}{d}}{p^d} = 1 - \prod_{i=0}^{d-1} \frac{p-i}{p}. \tag{13}$$

The term is built by subtracting the portion of required interactions (and their $d!$ permutations) form 1. Also here, we see that for fixed d, $\lim_{p\to\infty} \mathrm{E}_d(p) = 0$.

However, for fixed p and increasing degree, the quality of approximation reduces and the error $\mathrm{E}_d(p)$ reaches even 1 when d is larger than p (recall that the binomial coefficient is defined as being zero when $d > p$):

$$\mathrm{E}_d(p) = 1, \qquad \forall d > p.$$

This means

$$\lim_{p\to\infty} \mathrm{E}_d(p) = 0 \quad \text{and} \quad \lim_{d\to\infty} \mathrm{E}_d(p) = 1,$$

that is the error tends to 0 when d is fixed and p is increasing, but it approaches 1 when p is fixed and d is increasing.

In parts, this is an obvious result because dealing with a model with p markers, we can calculate the $(p+1)$th Hadamard power $\mathbf{G}^{\circ(p+1)}$, but there is no interaction between $p+1$ different loci. Also note that $\mathrm{E}_d(p)$ is a strictly monotonously increasing function for fixed p and increasing $d \leq p$ since

$$\prod_{i=0}^{d-1} \frac{p-i}{p} > \left(\prod_{i=0}^{d-1} \frac{p-i}{p}\right) \underbrace{\left(\frac{p-d}{p}\right)}_{<1} = \prod_{i=0}^{d} \frac{p-i}{p}$$

Moreover, considering the case $d = p$, we also receive

$$\lim_{p\to\infty} \mathrm{E}_p(p) = 1. \tag{14}$$

A proof of this statement can be found in the Appendix.

To illustrate this observation, let us consider a small example.

Example 1. Let us consider the case of five markers ($p = 5$) and the approximation of degree five ($d = 5$). Then

$$\mathrm{E}_5(5) = 1 - \prod_{i=0}^{4} \frac{5-i}{5} = \frac{601}{625} \approx 0.96 \tag{15}$$

Example 1 illustrates that the quality of the approximation can decrease quickly when the number of markers is very small.

4. Approximations for real genotypic data

Let us consider a data set of real genotypes. We use the marker data of a wheat data set published by Crossa et al. (2010) and also provided by the R package BGLR (Pérez and de Los Campos, 2014). For more information on the data set, see Crossa et al. (2010). We use a $\{\pm 1\}$ coding of the marker data, start with $p = 1$ and take the first marker of the data set to calculate $\mathbf{G}_1 = \mathbf{M}_1\mathbf{M}_1'$ and the corresponding $\mathbf{H}_1^{(1)}$. Note that for $d = 1$, $\mathbf{H}_p^{(1)} = \mathbf{G}_p$ for any p, meaning that the additive matrices are the same, or in other words, when drawing only one-element sets, it does not matter whether we draw with or without replacement.

We then subsequently increase the number of markers by adding the following columns of the marker matrix, and calculate the matrices $\mathbf{G}_p^{\circ d}$ and $\mathbf{H}_p^{(d)}$ for all $d \leq p$, that is the dth Hadamard power of $\mathbf{G}_p$ and the covariance matrix defined by interactions of d different markers out of the p markers. The correlation of the entries of $\mathbf{G}_p^{\circ d}$ and $\mathbf{H}_p^{(d)}$ is presented in Table 1 for $p \in \{1, \ldots, 25\}$ and $d \leq p$. We used the correlation of the entries as a similarity measure of the matrices because it is a simple criterion and independent of the data structure of a phenotype $\mathbf{y}$. Since $\mathbf{H}_p^{(d)}$ does not exist for $d > p$, no correlation is given for these cases. However, it is clear that the error of the approximation in the previously discussed sense is 100% for these cases. We see that for any fixed d and increasing p, the correlation of the entries of $\mathbf{G}_p^{\circ d}$ and $\mathbf{H}_p^{(d)}$ increases, but for any fixed p, the correlation reduces with increasing d.

An interesting aspect is that the correlation for $d = 2$ is directly equal to one, already for the case of $p = 2$. This is a result of using markers which only have two states coded as $\{\pm 1\}$. The matrix $\mathbf{G}_2^{\circ 2}$ is the sum of the matrices defined by the interactions $(\mathbf{M}_{\bullet,1}, \mathbf{M}_{\bullet,1})$, $(\mathbf{M}_{\bullet,1}, \mathbf{M}_{\bullet,2})$, $(\mathbf{M}_{\bullet,2}, \mathbf{M}_{\bullet,1})$ and $(\mathbf{M}_{\bullet,2}, \mathbf{M}_{\bullet,2})$. Contrarily,

Table 1
Correlation of the entries of $\mathbf{G}_p^{\circ d}$ and $\mathbf{H}_p^{(d)}$ for different values of p and d.

	d = 1,2	d = 3	d = 4	d = 5	d = 6	d = 7	d = 8	d = 9	d = 10	d = 11	d = 12	d = 13	d = 14	d = 15	d = 16	d = 17	d = 18	d = 19	d = 20	d = 21	d = 22	d = 23	d = 24	d = 25
p = 1	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 2	1	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 3	1	0.54	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 4	1	0.72	0.44	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 5	1	0.82	0.70	0.41	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 6	1	0.89	0.79	0.6	0.36	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 7	1	0.92	0.84	0.69	0.51	0.29	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 8	1	0.96	0.90	0.79	0.65	0.48	0.28	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 9	1	0.96	0.90	0.79	0.66	0.50	0.34	0.18	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 10	1	0.97	0.92	0.84	0.73	0.59	0.45	0.30	0.16	–	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 11	1	0.97	0.93	0.87	0.78	0.67	0.54	0.41	0.27	0.14	–	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 12	1	0.98	0.95	0.90	0.82	0.73	0.62	0.50	0.37	0.25	0.13	–	–	–	–	–	–	–	–	–	–	–	–	–
p = 13	1	0.98	0.95	0.90	0.83	0.75	0.66	0.56	0.44	0.31	0.20	0.10	–	–	–	–	–	–	–	–	–	–	–	–
p = 14	1	0.98	0.95	0.91	0.85	0.79	0.73	0.64	0.54	0.42	0.30	0.18	0.09	–	–	–	–	–	–	–	–	–	–	–
p = 15	1	0.99	0.96	0.92	0.88	0.83	0.77	0.70	0.62	0.52	0.40	0.28	0.18	0.09	–	–	–	–	–	–	–	–	–	–
p = 16	1	0.99	0.97	0.94	0.90	0.85	0.81	0.75	0.69	0.61	0.51	0.39	0.28	0.17	0.09	–	–	–	–	–	–	–	–	–
p = 17	1	0.99	0.97	0.94	0.90	0.87	0.83	0.79	0.74	0.68	0.60	0.50	0.38	0.26	0.16	0.08	–	–	–	–	–	–	–	–
p = 18	1	0.99	0.98	0.95	0.92	0.88	0.85	0.81	0.77	0.72	0.66	0.57	0.47	0.36	0.25	0.15	0.08	–	–	–	–	–	–	–
p = 19	1	0.99	0.98	0.95	0.92	0.90	0.87	0.85	0.82	0.78	0.73	0.67	0.59	0.49	0.37	0.25	0.15	0.07	–	–	–	–	–	–
p = 20	1	0.99	0.98	0.96	0.93	0.91	0.89	0.87	0.84	0.81	0.78	0.73	0.67	0.59	0.48	0.37	0.25	0.14	0.07	–	–	–	–	–
p = 21	1	0.99	0.98	0.96	0.94	0.93	0.91	0.89	0.87	0.85	0.82	0.79	0.74	0.68	0.61	0.51	0.39	0.26	0.15	0.07	–	–	–	–
p = 22	1	0.99	0.98	0.96	0.95	0.93	0.92	0.90	0.89	0.87	0.85	0.83	0.80	0.75	0.69	0.61	0.51	0.39	0.26	0.15	0.07	–	–	–
p = 23	1	0.99	0.98	0.97	0.95	0.94	0.93	0.92	0.91	0.89	0.88	0.86	0.83	0.80	0.75	0.69	0.61	0.51	0.39	0.26	0.14	0.06	–	–
p = 24	1	0.99	0.98	0.97	0.96	0.95	0.94	0.93	0.93	0.91	0.90	0.89	0.87	0.84	0.81	0.77	0.72	0.64	0.54	0.41	0.27	0.15	0.06	–
p = 25	1	0.99	0.98	0.97	0.96	0.95	0.95	0.94	0.93	0.92	0.91	0.90	0.88	0.86	0.83	0.80	0.76	0.70	0.62	0.52	0.40	0.26	0.15	0.06

$\mathbf{H}_2^{(2)}$ corresponds only to the interaction $(\mathbf{M}_{\bullet,1}, \mathbf{M}_{\bullet,2})$. Thus, here $\mathbf{G}_2^{\circ 2}$ is twice $\mathbf{H}_2^{(2)}$ plus the matrices corresponding to the interactions $(\mathbf{M}_{\bullet,1}, \mathbf{M}_{\bullet,1})$ and $(\mathbf{M}_{\bullet,2}, \mathbf{M}_{\bullet,2})$. However, since we use the $\{\pm 1\}$ coding – and the interactions are modeled by the squared marker values – each of the two latter matrices corresponds to a relationship matrix defined by a marker which is equal to 1 for each individual. Consequently, we only add a constant matrix, which does not have an impact on the correlation between $\mathbf{G}_2^{\circ 2}$ and $\mathbf{H}_2^{(2)}$.

Also if we consider the cases $d = p$, we see that the correlation of the matrices tends to 0 for increasing p, which has already been stated by Eq. (14).

5. Limit considerations of models with higher degree interactions

In the following, we would like to investigate empirically whether the decreasing quality of the approximation for higher degree interactions has an impact on limit considerations.

Limit Problem 1. We would like to build a model that takes all interactions of p different loci into account. We assume for this limit model that the variance component σ_β^2 remains the same for any degree d, that is any interaction effect of any degree comes from the same distribution. We can formulate the model which we are interested in as

$$\sigma_\beta^2 \left(\mathbf{H}^{(1)} + \mathbf{H}^{(2)} + \mathbf{H}^{(3)} + \cdots \mathbf{H}^{(p)}\right) \tag{16}$$

Since it is computationally demanding to calculate the matrices $\mathbf{H}^{(d)}$ for higher degree interaction, we are interested in an approximation using Hadamard powers $\mathbf{G}^{\circ d}$. As discussed above, the matrix $\mathbf{G}^{\circ d}$ counts each interaction which we aim to model in $\mathbf{H}^{(d)}$ $d!$ times. Moreover, additional interactions are included in $\mathbf{G}^{\circ d}$ in which we are not interested. To give an equal weight to any interaction which we would like to model, we have to divide each $\mathbf{G}^{\circ d}$ by $d!$ to guarantee that the weights are adapted between degrees. Without this adjustment, we would model the interactions of degree two twice giving them twice the weight of the additive effects. Analogously, the interactions of degree three would be modeled six times, giving each of them six times the weight of an additive effect.

An approximation of our desired relationship model can thus be given by

$$\sigma_\beta^2 \left(\mathbf{G} + \frac{1}{2!}\mathbf{G}^{\circ 2} + \frac{1}{3!}\mathbf{G}^{\circ 3} + \cdots \frac{1}{p!}\mathbf{G}^{\circ p}\right) \tag{17}$$

where we include the inverse factorial to scale the matrices relatively to each other. Since each entry in Eq. (17) follows the exponential power series, it can be approximated for large p by

$$\sigma_\beta^2 \left(\exp(\mathbf{G}) - \mathbf{1}_{n\times n}\right). \tag{18}$$

Please recall that the operations are here meant entry-wise. In particular, the exponential function refers to the entry-wise exponential (and not the matrix exponential): $(\exp(\mathbf{G}))_{i,j} := \exp(\mathbf{G}_{i,j})$. Moreover, note here that this limit is not identical to the Gaussian kernel in a reproducing kernel Hilbert space approach, since the exponential function is not applied to the squared Euclidean distance but to the entries of $\mathbf{G}$ (which is slightly different from a limit consideration leading to the Gaussian kernel and presented by Jiang and Reif (2015)).

A question is how good the approximation of the covariance model which we actually would like to model (Eq. (16)) by Eq. (18) is. Although the presence of the inverse factorials, which give the weights to the Hadamard powers of $\mathbf{G}$ in Eq. (17), suggests that the influence of higher degree interactions will quickly vanish, a general theoretical consideration is difficult since the quality of approximation also depends on how fast the entries of $\mathbf{G}^p$ grow.

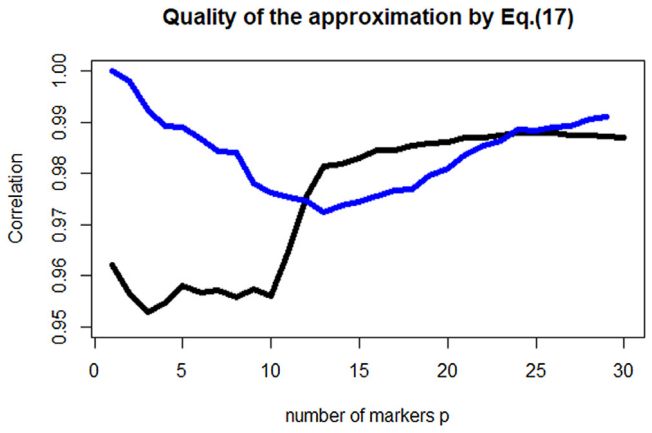

Fig. 1. Correlation of the entries of the matrices defined by Eqs. (16) and (18) for increasing p for the wheat data set (blue line) and the maize data set (black line). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

For this reason, we use the relationship matrices calculated for the wheat data set (Table 1) as well as a $\{-1, 0, 1\}$ coded maize data (for more details on the data see the section Data at the end of the manuscript) and consider the correlation of the matrices defined by Eqs. (16) and (18) for increasing p. The results are presented in Fig. 1. We see for the wheat data (blue line) an initially high correlation which is a result of only using one marker. Since we have only one marker with the two possible values $\{\pm 1\}$, each entry of $\mathbf{G}_1$ has only two possibilities. Applying the non-linear exponential function still gives a matrix with only two values which is perfectly correlated with $\mathbf{G}_1$. This high correlation is then reduced when a second marker is introduced and the correlation keeps decreasing until the increase in p improves the approximation sufficiently to push the correlation again towards 1. For the maize data (black line), the correlation starts – since we are dealing with markers with three states – on a lower level but increases quickly towards 1.

Limit Problem 2. Let us assume that we would like to use a model in which the variance of the higher degree interaction increases by $d!$

$$\sigma_\beta^2 \left(\mathbf{H}^{(1)} + 2!\mathbf{H}^{(2)} + 3!\mathbf{H}^{(3)} + \cdots p!\mathbf{H}^{(p)}\right). \tag{19}$$

It should be mentioned here that epistatic effects are usually defined as deviations from the fit defined by lower degree interactions. This concept would translate to the assumption that the variance components decrease with increasing d. However, Eq. (19) is a valid covariance model and we would like to investigate the effect of these increasing weights on the overall approximation. A reason for defining an increasing variance could be to give the higher degree interactions more flexibility to capture some important interaction terms.

We may approximate Eq. (19) by

$$\sigma_\beta^2 \left(\mathbf{G} + \mathbf{G}^{\circ 2} + \mathbf{G}^{\circ 3} + \cdots \mathbf{G}^{\circ p}\right) \tag{20}$$

which converges for increasing p to

$$\frac{1}{(1-\mathbf{G})} - \mathbf{1}_{n\times n} \tag{21}$$

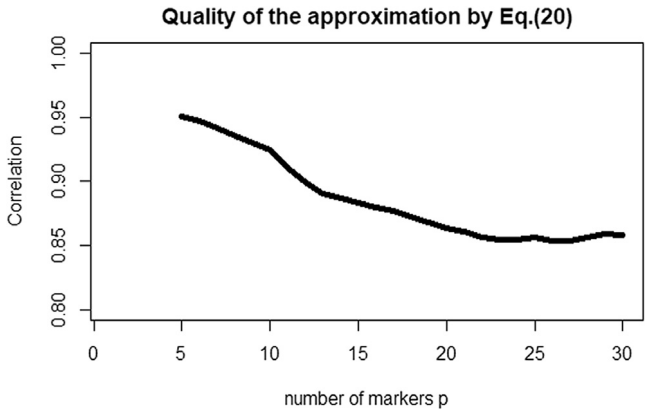

Fig. 2. Correlation of the entries of the matrices defined by Eqs. (19) and (21) for increasing p and the maize data (coded as $\{-1, 0, 1\}/\sqrt{p}$).

iff $|G_{i,j}| < 1\ \forall i, j$. The latter condition of the entries having absolute values smaller than 1 is essential, since otherwise the series of Eq. (20) will not converge. Recall again that all operations are meant entry-wise.

This condition of not any absolute entry being larger than or equal to 1 is for instance given when we are dealing with $\{-1, 0, 1\}$ coded data, the marker data is divided by the square root of p (which is equal to dividing $\mathbf{G}$ by p), and if none of the lines is completely homozygous. The wheat data is not appropriate for this limit, since it has only two values and the entries on the diagonal would be equal to 1 (when dividing the markers by $\sqrt{p}$). However, we consider the behavior of the correlation for the maize data (with dividing the marker values by $\sqrt{p}$).

Fig. 2 illustrates that – since the weights of the higher degree interactions are not reduced in Eq. (20) – the accumulated error across the different degrees d matters more than in Limit Problem 1. The correlation of the entries of the matrices defined by Eqs. (19) and (21) reduces, and a reversion of this trend cannot be observed up to $p = 30$, which was the maximal value for which we were able to calculate $\mathbf{H}^{(p)}$ with our approach. Note here that for values of p below five, some of the included lines were completely homozygous, which leads to a situation of the maximal value of the additive relationship matrix being 1 and thus Eq. (21) not being defined. Therefore, no correlation is given for these points.

6. Summary and outlook

We gave an explicit formula to quantify the error when approximating a model with interactions between different loci by Hadamard powers of the additive genomic relationship (Eq. (13)). The criterion used to quantify the quality of the approximation also struggles with the problem of how to add a new column of markers when p increases, but gives a simple and well-defined equation.

We illustrated that when the number of markers p is fixed and d increases, the quality of the approximation of $\mathbf{H}^{(d)}$ by $\mathbf{G}^{\circ d}$ decreases. For limit considerations such as Limit Problem 1, where the impact of higher degree interactions reduces quickly, this reduced quality does not have a big impact on the quality of approximation of the overall limit. However – as illustrated by Limit Problem 2 – for models in which the weight is not reduced fast enough with increasing degree d, the overall limit can have a lower (but still high) correlation with what is supposed to be approximated.

Due to the computational restrictions, we were not able to calculate all $\mathbf{H}^{(d)}$ with $d \leq p$ for p larger than 30. Thus, we cannot judge the behavior of our empirical considerations of Limit Problems 1 and 2 above 30 markers. An interesting theoretical problem – which would also allow to investigate the behavior of limits for larger values of p – would be to find a concise equation generalizing Eq. (4) to any degree d.

Data

As described above, the wheat data has been published by Crossa et al. (2010) and is also provided by the R package BGLR (Pérez and de Los Campos, 2014). The maize data was provided by the same publication (Crossa et al., 2010). We used the file dataCorn_SS_asi.RData which is available in File S1 of following link

https://www.genetics.org/content/186/2/713.supplemental

We reduced the set to the 101 lines which had at least one heterozygous (0) marker within the first five markers. Since calculating $\mathbf{H}^{(d)}$ is computationally demanding, we restricted us to this subset for which Eq. (21) is already defined for $p = 5$. This reduced data set was used for Limit Problems 1 and 2.

Acknowledgments

We are thankful for the financial support provided by CIMMYT, CGIAR CRP WHEAT, the Bill & Melinda Gates Foundation, as well as the USAID projects (Cornell University and Kansas State University) that generated the CIMMYT wheat data analyzed in this study. We acknowledge the financial support provided by the Foundation for Research Levy on Agricultural Products (FFL) and the Agricultural Agreement Research Fund (JA) in Norway through NFR grant 267806. Moreover, we thank two anonymous referees, especially the one who pointed out an important error in the Appendix.

Appendix A. Extensions of Eq. (4) to $d \in \{3, 4\}$ and illustration of the general problem

As pointed out earlier, the difference between $\mathbf{G}^{\circ d}$ and $\mathbf{H}^{(d)}$ is represented by the difference in the sets of possible events when drawing d times from $\{1, \ldots, p\}$ either with or without replacement. For $d = 2$, we can simply use the matrix corresponding to the set of interactions $\{1, \ldots, p\} \times \{1, \ldots, p\}$, remove the covariance matrix coming from the tuples (i, i) and divide the remaining matrix by 2 to account for not considering the order of the draws, which means (i, j) is considered to be equal to (j, i). The matrix $\mathbf{G}^{\circ 2}$ corresponds to $\{1, \ldots, p\} \times \{1, \ldots, p\}$ and the matrix $(\mathbf{M}^{\circ 2})(\mathbf{M}^{\circ 2})'$ represents the interactions $(i, i)_{i=1,\ldots,p}$.

Before, we go to the cases of $d \in \{3, 4\}$, recall that we are looking for a concise formula, that is one that can be easily calculated using Hadamard products of $\mathbf{M}$ and $\mathbf{G}$. It is obvious that there are equations which allow to calculate $\mathbf{H}^{(d)}$, for instance the straight-forward approach used in this manuscript, but we are looking for a formula allowing the use of Hadamard products and simplifying the computation.

Let us now consider the case of $d = 3$. Analogously to the case of $d = 2$, we identify $\mathbf{G}^{\circ 3}$ with the set of interactions described by $\{1, \ldots, p\} \times \{1, \ldots, p\} \times \{1, \ldots, p\}$. We have to subtract the matrix corresponding to $\{(i, i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}$ and its permutations. This can be represented by the matrix $3\left((\mathbf{M}^{\circ 2})(\mathbf{M}^{\circ 2})'\right) \circ \mathbf{G}$. However, this matrix also includes the covariance generated by $\{(i, i, i)_{i=1,\ldots,p}\}$ and each of these tuples only occurs once, not three times. Thus, we have to correct what we had subtracted by adding twice the matrix $(\mathbf{M}^{\circ 3})(\mathbf{M}^{\circ 3})'$. This procedure is similar to

the "principle of inclusion and exclusion" known from basic set theory and gives the following identity

$$\mathbf{H}^{(3)} = \frac{1}{6}\left(\mathbf{G}^{\circ 3} - 3\left((\mathbf{M}^{\circ 2})(\mathbf{M}^{\circ 2})'\right) \circ \mathbf{G} + 2(\mathbf{M}^{\circ 3})(\mathbf{M}^{\circ 3})'\right).$$

Analogously, a formula for degree $d = 4$ can be derived. However, we have to consider here the sets

$$\{1, \ldots, p\}^{\times 4}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 2}$$

$$\{(i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}$$

$$\{(i,i,i,i)_{i=1,\ldots,p}\}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j)_{j=1,\ldots,p}\}$$

with their corresponding permutations. Moreover, an important point is to be aware of the fact that the sets are in parts subsets of others. Thus, we obtain

$$\mathbf{H}^{(4)} = \frac{1}{24}\Big(\mathbf{G}^{\circ 4} - 6(\mathbf{M}^{\circ 2})(\mathbf{M}^{\circ 2})' \circ \mathbf{G}^{\circ 2} + 3\left((\mathbf{M}^{\circ 2})(\mathbf{M}^{\circ 2})'\right)^{\circ 2} + 8(\mathbf{M}^{\circ 3})(\mathbf{M}^{\circ 3})' \circ \mathbf{G} - 6(\mathbf{M}^{\circ 4})(\mathbf{M}^{\circ 4})'\Big).$$

Each summand on the right hand side corresponds to one of the sets mentioned above.

For $d = 5$, the following sets would have to be considered:

$$\{1, \ldots, p\}^{\times 5}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 3}$$

$$\{(i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 2}$$

$$\{(i,i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}$$

$$\{(i,i,i,i,i)_{i=1,\ldots,p}\}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j)_{j=1,\ldots,p}\} \times \{1, \ldots, p\}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j,j)_{j=1,\ldots,p}\}$$

And for $d = 6$:

$$\{1, \ldots, p\}^{\times 6}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 4}$$

$$\{(i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 3}$$

$$\{(i,i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 2}$$

$$\{(i,i,i,i)_{i=1,\ldots,p}\} \times \{(j,j)_{j=1,\ldots,p}\}$$

$$\{(i,i,i,i,i)_{i=1,\ldots,p}\} \times \{1, \ldots, p\}$$

$$\{(i,i,i,i,i,i)_{i=1,\ldots,p}\}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j)_{j=1,\ldots,p}\} \times \{1, \ldots, p\}^{\times 2}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j,j)_{j=1,\ldots,p}\} \times \{1, \ldots, p\}$$

$$\{(i,i,i)_{i=1,\ldots,p}\} \times \{(j,j,j)_{j=1,\ldots,p}\}$$

$$\{(i,i)_{i=1,\ldots,p}\} \times \{(j,j)_{j=1,\ldots,p}\} \times \{(k,k)_{k=1,\ldots,p}\}$$

It is not straightforward to extend this procedure to higher degrees and in an automated way. An approach might be to try to find a recursive formula that allows to calculate $\mathbf{H}^{(d)}$ from $\{\mathbf{H}^{(i)}\}_{i=1,\ldots,d-1}$.

Appendix B. Challenging Eq. (11)

B.1. Alternating entries

For instance, consider the sequence $(\mathbf{M}_p)_{p\in\mathbb{N}}$ defined by the pth column being $\mathbf{M}_{\bullet,p} := \begin{pmatrix}1\\1\end{pmatrix}$ when p is odd or $\mathbf{M}_{\bullet,p} := \begin{pmatrix}-1\\1\end{pmatrix}$ when p is even. Then $\mathbf{G}_p$ will alternate between

$$\mathbf{G}_p = \begin{pmatrix}p & 0\\0 & p\end{pmatrix} \text{ and } \mathbf{G}_{p+1} = \begin{pmatrix}p+1 & 1\\1 & p+1\end{pmatrix}$$

In particular this means that the diagonal will increase with p and "tend to ∞", but the off-diagonal elements will alternate. Here, the limit of $\tilde{\mathbf{H}}$ is not defined (even if "∞" is considered as a limit), that is Eq. (11) does not make sense for this example. Note here that one could argue that the limit $\mathbf{G}_p/p$ would be well-defined and that we have to use this alternative definition of the genomic relationship matrix. However, we can define a more complicated example for which $\mathbf{G}_p/p$ will neither converge.

B.2. An example for which $\mathbf{G}_p/p$ does not converge and thus Eq. (11) is not defined

Start with $\mathbf{M}_1 := \begin{pmatrix}1\\1\end{pmatrix}$ and add a column $\begin{pmatrix}-1\\1\end{pmatrix}$ for $\mathbf{M}_2$. Add two columns $\begin{pmatrix}1\\1\end{pmatrix}$ followed by four columns of $\begin{pmatrix}-1\\1\end{pmatrix}$. Whenever, the sign of the entry is switched, add as many columns with the respective sign as you already have when switching the sign. For instance $\mathbf{M}_8$ would be

$$\mathbf{M}_8 := \begin{pmatrix}1 & -1 & 1 & 1 & -1 & -1 & -1 & -1\\1 & 1 & 1 & 1 & 1 & 1 & 1 & 1\end{pmatrix}.$$

We would like to thank an anonymous reviewer here who pointed out that – based on the above definition – we can write

$$\mathbf{G}_{2^k} = \begin{pmatrix}2^k & 1+\sum_{i=1}^{k}\frac{(-2)^i}{2}\\1+\sum_{i=1}^{k}\frac{(-2)^i}{2} & 2^k\end{pmatrix} = \begin{pmatrix}2^k & \frac{(-2)^k+2}{3}\\\frac{(-2)^k+2}{3} & 2^k\end{pmatrix}$$

Consequently, $\mathbf{G}_p/p$ has a subsequence converging to

$$\begin{pmatrix}1 & 1/3\\1/3 & 1\end{pmatrix}$$

and another subsequence converging to

$$\begin{pmatrix}1 & -1/3\\-1/3 & 1\end{pmatrix}.$$

Thus, $\mathbf{G}_p/p$ does not converge.

The examples described above aimed at illustrating the problem of possibly alternating entries. Another important situation to consider is the convergence of the entries of $\mathbf{G}_p$ to zero.

B.3. Zero as a limit

Convergence to zero can also cause trouble for Eq. (11). Consider for instance

$$\tilde{\mathbf{H}}_p = \begin{pmatrix}2/p & 1/p\\1/p & 2/p\end{pmatrix} \text{ and } \mathbf{H}_p = \begin{pmatrix}2/p & 2/p\\2/p & 2/p\end{pmatrix}.$$

Both matrices have the same limit, which means Eq. (11) is fulfilled with $c = 1$, but in $\tilde{\mathbf{H}}_p$ the variances will always have the double weight of the covariances, whereas all entries are identical in $\mathbf{H}_p$. Thus, the quality of approximation when approximating one by the other should remain on the same level — independent of p.

B.4. Tending to infinity

Another question would be whether we allow the entries of the matrices to tend to infinity as a limit in Eq. (11), which would cause additional problems for the definition of "an improvement of the quality of the approximation" by Eq. (11). Let us for instance consider the hypothetical example of

$$\tilde{\mathbf{H}}_p = \begin{pmatrix} p+2 & p-2 \\ p-2 & p+2 \end{pmatrix} \text{ and } \mathbf{H}_p = \begin{pmatrix} p & p \\ p & p \end{pmatrix}.$$

Both matrices converge to ∞ and thus they – considering ∞ as a limit – have the same limit. However, they will remain equally distant for – for instance – the Euclidean metric.

One approach could be to consider the difference for each entry of $\tilde{\mathbf{H}}_p$ and $\mathbf{H}_p$ relatively to the respective entry of $\mathbf{H}_p$. However, this will also lead to problems for Eq. (11), if an entry converges to or is equal to 0.

Appendix C. Proof of Eq. (14)

We would like to thank here again an anonymous reviewer who pointed at the importance of Eq. (14) and also provided this proof: According to the definition,

$$\mathrm{E}_p(p) := 1 - \prod_{i=0}^{p-1} \frac{p-i}{p} = 1 - \frac{p!}{p^p}.$$

We have to show that $f(p) := \frac{p!}{p^p}$ tends to 0 when p increases. Consider

$$f(p+1) = \frac{(p+1)!}{(p+1)^{(p+1)}} = \frac{p!}{(p+1)^p} = f(p) \cdot \underbrace{\frac{p^p}{(p+1)^p}}_{<1}.$$

This shows that $f(p)$ is monotonously decreasing and thus converges to a number since it has the lower bound 0. Moreover, $\frac{p^p}{(p+1)^p}$ tends to $1/e$. Since both factors have well-defined limits,

$$\lim_{p\to\infty} f(p) = \lim_{p\to\infty} f(p+1) = \lim_{p\to\infty} f(p) \cdot \frac{1}{e}$$

which can only be satisfied if

$$\lim_{p\to\infty} f(p) = 0. \qquad \square$$

References

de los Campos, G., Gianola, D., Rosa, G., 2009. Reproducing kernel Hilbert spaces regression: a general framework for genetic evaluation. J. Anim. Sci. 87, 1883–1887. http://dx.doi.org/10.2527/jas.2008-1259.

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J.L., Makumbi, D., Singh, R.P., Dreisigacker, S., Yan, J., Arief, V., Banziger, M., Braun, H.J., 2010. Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics 186, 713–724. http://dx.doi.org/10.1534/genetics.110.118521.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., Burgueño, J., González-Camacho, J.M., Pérez-Elizalde, S., Beyene, Y., et al., 2017. Genomic selection in plant breeding: methods, models, and perspectives. Trends Plant Sci. 22 (11), 961–975.

Falconer, D.S., Mackay, T.F.C., 1996. Introduction to Quantitative Genetics. Pearson Education, London.

Gianola, D., Gustavo, A., Hill, W.G., Manfredi, E., Fernando, R.L., 2009. Additive genetic variability and the bayesian alphabet. Genetics.

Hayes, B.J., Bowman, P.J., Chamberlain, A., Goddard, M., 2009. Invited review: Genomic selection in dairy cattle: Progress and challenges. J. Dairy Sci. 92 (2), 433–443.

He, D., Parida, L., 2016. Does encoding matter? a novel view on the quantitative genetic trait prediction problem. BMC Bioinform. 17 (9), 272.

Jannink, J.-L., Lorenz, A.J., Iwata, H., 2010. Genomic selection in plant breeding: from theory to practice. Brief. Funct. Genom. 9 (2), 166–177.

Jiang, Y., Reif, J.C., 2015. Modeling epistasis in genomic selection. Genetics 201, 759–768. http://dx.doi.org/10.1534/genetics.115.177907.

Martini, J.W.R., Gao, N., Cardoso, D.F., Wimmer, V., Erbe, M., Cantet, R.J.C., Simianer, H., 2017. Genomic prediction with epistasis models: on the marker-coding-dependent performance of the extended GBLUP and properties of the categorical epistasis model (CE). BMC Bioinformatics 18, 3. http://dx.doi.org/10.1186/s12859-016-1439-1.

Martini, J.W.R., Rosales, F., Ha, N.-T., Heise, J., Wimmer, V., Kneib, T., 2019. Lost in translation: On the problem of data coding in penalized whole genome regression with interactions. G3: Genes Genomes Genet. g3–200961.

Martini, J.W.R., Wimmer, V., Erbe, M., Simianer, H., 2016. Epistasis and covariance: how gene interaction translates into genomic relationship. Theor. Appl. Genet. 129, 963–976. http://dx.doi.org/10.1007/s00122-016-2675-5.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E., 2001. Prediction of total genetic value using genome-wide dense marker maps. Genetics 157, 1819–1829, URL http://www.genetics.org/content/157/4/1819.

Meuwissen, T.H.E., Hayes, B.J., Goddard, M., 2016. Genomic selection: A paradigm shift in animal breeding. Anim. Front. 6 (1), 6–14.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., Simianer, H., Hoeschele, I., 2011. Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. Genetics 188, 695–708. http://dx.doi.org/10.1534/genetics.111.128694.

Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., Mackay, T.F.C., 2015. Accounting for genetic architecture improves sequence based genomic prediction for a *Drosophila* fitness trait. PLOS ONE 10, e0126880. http://dx.doi.org/10.1371/journal.pone.0126880.

Pérez, P., de Los Campos, G., 2014. Genome-wide regression and prediction with the BGLR statistical package. Genetics 198 (2), 483–495.

Strandén, I., Christensen, O.F., 2011. Allele coding in genomic evaluation. Genet. Sel. Evol. 43 (1), 25.

Su, G., Christensen, O.F., Ostersen, T., Henryon, M., Lund, M.S., 2012. Estimating additive and non-additive genetic variances and predicting genetic merits using genome-wide dense single nucleotide polymorphism markers. PLoS One 7 (9), e45293.

VanRaden, P.M., 2008. Efficient methods to compute genomic predictions. J. Dairy Sci. 91 (11), 4414–4423.

Xu, S., 2013. Mapping quantitative trait loci by controlling polygenic background effects. Genetics 195 (4), 1209–1222.