Received: 30 October 2019 | Accepted: 18 April 2020 | Published online: 23 July 2020

DOI: 10.1002/tpg2.20033

The Plant Genome

ORIGINAL RESEARCH

On Hadamard and Kronecker products in covariance structures for genotype × environment interaction

Johannes W. R. Martini[1] | **Jose Crossa**[1] | **Fernando H. Toledo**[1] | **Jaime Cuevas**[2]

[1] International Maize and Wheat Improvement Center (CIMMYT), Km. 45, El Batán 56237 Texcoco, Mexico

[2] Universidad de Quintana Roo, Del Bosque, 77019 Chetumal, Q.R., Mexico

Correspondence
Johannes W. R. Martini, International Maize and Wheat Improvement Center (CIMMYT), Km. 45, El Batán 56237 Texcoco, Mexico.
Email: j.martini@cgiar.org

Funding information
Bill & the Melinda Gates Foundation; Foundation for Research Levy on Agricultural Products; Agricultural Agreement Research Fund, Grant/Award Number: 267806

Abstract

When including genotype × environment interactions (G × E) in genomic prediction models, Hadamard or Kronecker products have been used to model the covariance structure of interactions. The relation between these two types of modeling has not been made clear in genomic prediction literature. Here, we demonstrate that a certain model based on a Hadamard formulation and another using the Kronecker product lead to exactly the same statistical model. Moreover, we illustrate how a multiplication of entries of covariance matrices is related to modeling locus × environmental-variable interactions explicitly. Finally, we use a wheat and a maize data set to illustrate that the environmental covariance **E** can be specified easily, also if no information on environmental variables – such as temperature or precipitation – is available. Given that lines have been tested in different environments, the corresponding environmental covariance can simply be estimated from the training set as phenotypic covariance between environments. To achieve a high level of increase in predictive ability, the environmental covariance has to be defined appropriately and records on the performance of the lines of the test set under different environmental conditions have to be included in the training set.

1 | INTRODUCTION

Multi-environment trials (MET) with which lines are evaluated under several environmental conditions are fundamental in plant breeding. The comparison of their yield-stability and their genotype × environment interaction (G × E) allows the breeder to select the best genotypes for specific environments or the lines with the most stable yield performance across environments. Early approaches for the analysis of G × E used fixed effects models ignoring that the continuous degree of similarity of environmental conditions at different locations may potentially be incorporated in the statistical model. Similarly, relationship of the lines, as well as spatial trends in the field have not been considered in these early approaches (Cornelius & Crossa, 1999; Cornelius, Crossa, & Seyedsadr, 1996; Crossa & Cornelius, 1997; Cornelius et al., 2001; Crossa et al., 2001). A restriction at that time was the limited availability of appropriate statistical software.

One of the frameworks that easily allows including covariance structures between different random effects are mixed linear models. Mixed linear models and the

Abbreviations: BLUP, best linear unbiased prediction; E1, environment 1; E2, environment 2; E3, environment 3; E4, environment 4; G × E, genotype by environment interaction; GBLUP, genomic best linear unbiased prediction; MET, multi-environment trials.

corresponding best linear unbiased prediction (BLUP) – classically treated by Henderson (1975) – modify the least square regression approach to a penalized regression and provide a framework to incorporate correlations between sites, years, and plots in the field, as well as genetic relation between relatives.

Crossa et al. (2006) illustrated how genetic effects can be modeled as additive and additive × environment interaction using a factor analytic model across environments and the additive pedigree relationship **A**. The authors showed that modeling G × E increases the precision of the predictions. Burgueño, Crossa, Cornelius, Trethowan, McLaren, and Krishnamachari (2007) split the total genetic effects into pedigree based additive and additive × additive effects and modeled the additive × environment interaction and the additive × additive × environment interaction.

With the advances in genotyping methods, dense marker based genomic prediction (Meuwissen, Hayes, & Goddard, 2001) and the selection for predicted genetic values (genomic selection) has become a central component of breeding strategies (Crossa et al., 2011; Crossa et al., 2017; Hayes, Bowman, Chamberlain, & Goddard, 2009; Jannink, Lorenz, & Iwata, 2010). After comparing different types of (genomic) relationship (e.g. Crossa et al., 2010; de los Campos, Gianola, & Rosa, 2009; de los Campos, Gianola, Rosa, Weigel, & Crossa, 2010), a second wave of research has addressed G × E interaction in combination with the genomic relationship matrix **G** (e.g. Burgueño, de los Campos, Weigel, & Crossa, 2012; Lopez-Cruz et al., 2015; Yao et al., 2017; Gonzales-Barrios, Diaz-Garcia, & Gutierrez, 2019).

In some approaches, the covariance structure used as input for the G × E interaction has been modeled using Kronecker products (Cuevas, Crossa, Montesinos-López, Burgueño, Pérez-Rodríguez, & de los Campos, 2017), others used Hadamard products (Acosta-Pech et al., 2017; Basnet et al., 2019; Jarquin et al., 2014; Perez-Rodriguez et al., 2015; Perez-Rodriguez et al., 2017; Sukumaran et al., 2017). Moreover, Kronecker products have been used to model the interaction between the parental genomes of hybrids to capture specific combining ability effects (Acosta-Pech et al., 2017; Basnet et al., 2019). When addressing epistasis, that is genetic interaction, Hadamard products of the additive genomic relationship have mainly been used (e.g. Gao et al., 2017; Jiang & Reif, 2015; Martini, Toledo, & Crossa, 2020; Martini, Wimmer, Erbe, & Simianer, 2016; Varona, Legarra, Toro, & Vitezica, 2018; Vitezica, Legarra, Toro, & Varona, 2017).

As described, both mathematical operations – the Hadamard and the Kronecker product – have been used to model the covariance of interactions. However, it has not been pointed out what the relation between both types of modeling is and whether and how these approaches differ. In this study, we give a theoretical proof that shows that both methods lead to exactly the same covariance model when used with appropriate design matrices reflecting the arrangement of the phenotypic data. Moreover, we illustrate how multiplying the entries of covariance matrices relates to modeling the interaction between underlying variables, such as markers, temperature or precipitation, explicitly. Finally, we use two publicly available data sets to illustrate that G × E models can improve the prediction accuracy also if no information on the environmental conditions is available. Given that the training set includes a sufficient overlap of lines across environments, estimating the phenotypic covariance from the training set is sufficient to specify the mixed model.

Core Ideas

- We compare the structure of covariance models for genotype × environment interactions defined by Hadamard or by Kronecker products
- We highlight the identity of modeling marker × environmental-variable interaction and multiplying covariances
- We present a heuristic approach for defining the environmental covariance

2 | THEORETICAL BACKGROUND

We will shortly recapitulate the Hadamard, the Kronecker, and the Khatri-Rao product and illustrate how they can be used to model G × E covariance.

2.1 | Recapitulation of the Hadamard product

The Hadamard product $\circ$ is defined for two matrices of equal size. Let **A** and **B** be two matrices of size $n \times m$. Then the Hadamard product $\mathbf{A} \circ \mathbf{B}$ is defined by

$$(\mathbf{A} \circ \mathbf{B})_{i,j} := (\mathbf{A})_{i,j} \cdot (\mathbf{B})_{i,j}$$

where $i = 1,2,\ldots,n$ indicates the row and $j = 1,2,\ldots,m$ the column. The symbol := means that we define the left hand-side by the right-hand side. Moreover, the symbol $\cdot$ denotes the ordinary multiplication of the real numbers. In words, **A** and **B** are multiplied entry-wise and $\mathbf{A} \circ \mathbf{B}$ is also of size $n \times m$. Note here that we will be dealing with covariance matrices which are by definition square, that is of size $n \times n$. Moreover note that $\mathbf{A} \circ \mathbf{B} = \mathbf{B} \circ \mathbf{A}$.

2.2 | Recapitulation of the Kronecker product

The Kronecker product $\otimes$ is defined for any two matrices $\mathbf{A}$ and $\mathbf{B}$, not necessarily of same size. For $\mathbf{A}$ of size $n \times m$ and $\mathbf{B}$ of size $k \times l$,

$$\mathbf{A} \otimes \mathbf{B} := \begin{pmatrix} a_{1,1}\mathbf{B} & a_{1,2}\mathbf{B} & \dots & a_{1,(m-1)}\mathbf{B} & a_{1,m}\mathbf{B} \\ a_{2,1}\mathbf{B} & a_{2,2}\mathbf{B} & & a_{2,(m-1)}\mathbf{B} & a_{2,m}\mathbf{B} \\ \vdots & & \ddots & \vdots & \\ a_{n,1}\mathbf{B} & a_{n,2}\mathbf{B} & \cdots & a_{n,(m-1)}\mathbf{B} & a_{n,m}\mathbf{B} \end{pmatrix}$$

Each block $a_{i,j}\mathbf{B}$ represents a matrix of size $k \times l$, which means $\mathbf{A} \otimes \mathbf{B}$ is of size $(n \cdot k) \times (m \cdot l)$. For square matrices, their Kronecker product will be square. Note that $\mathbf{A} \otimes \mathbf{B} \neq \mathbf{B} \otimes \mathbf{A}$ since the entries of the products will have a different order.

2.3 | Recapitulation of the Khatri-Rao product

We recapitulate a third type of product, the column-wise Kronecker or the Khatri-Rao product. The Khatri-Rao product $*$ is defined for any two matrices with the same number of columns. For two matrices $\mathbf{A}$ of size $n \times m$ and $\mathbf{B}$ of size $k \times m$, the i-th column of $(\mathbf{A} * \mathbf{B})$ is defined by

$$(\mathbf{A} * \mathbf{B})_{\bullet,i} := (\mathbf{A})_{\bullet,i} \otimes (\mathbf{B})_{\bullet,i}$$

Consequently, $(\mathbf{A} * \mathbf{B})$ will be of size $(n \cdot k) \times m$. In the case that $\mathbf{A}$ and $\mathbf{B}$ are column vectors, the Khatri-Rao product is identical to the Kronecker product. Note that $\mathbf{A} * \mathbf{B} \neq \mathbf{B} * \mathbf{A}$ since the entries in each column will have a different order.

For more information see for instance Kolda and Bader (2009). In this manuscript, the Khatri-Rao product will only be relevant for the precise definition of a design matrix in the second theoretical result.

2.4 | Hadamard and Kronecker products for modeling G × E covariance

Let us assume that we are dealing with three plant lines and phenotypic records in two different environments. The phenotypic data is given by the 6×1 column vector

$$\mathbf{y} = \left(y_{1,1}, y_{1,2}, y_{2,1}, y_{2,2}, y_{3,1}, y_{3,2}\right)^t$$

where $y_{i,j}$ is the phenotype of the i-th line in the j-th environment. We use the statistical model

$$\mathbf{y} = \mu \mathbf{1}_6 + \mathbf{Z}_1\mathbf{g} + \mathbf{Z}_2\mathbf{e} + \mathbf{Z}_3(\mathbf{ge}) + \boldsymbol{\varepsilon} \tag{1}$$

where μ denotes a fixed effect, $\mathbf{1}_6$ is a 6×1 vector with each entry equal to 1, $\mathbf{g} \sim \mathbf{N}(0, \sigma_g^2\, \mathbf{G})$ the 3×1 genetic effects with a covariance (genomic relationship) matrix $\mathbf{G}$, and $\mathbf{e} \sim \mathbf{N}(0, \sigma_e^2\, \mathbf{E})$ a 2×1 random environmental effect with a covariance $\mathbf{E}$. The matrix $\mathbf{E}$ can be the identity matrix but also a structured covariance obtained for instance from phenotypic correlations across environments or environmental variables (e.g. Perez-Rodriguez et al., 2015). Moreover, the error is assumed to be $\boldsymbol{\varepsilon} \sim \mathbf{N}(0, \sigma_\varepsilon^2\, \mathbf{I})$ with identity matrix $\mathbf{I}$, meaning that the error terms are independent and identically distributed. The design matrices $\mathbf{Z}_1$, $\mathbf{Z}_2$ and $\mathbf{Z}_3$ are given by

$$\mathbf{Z}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \end{pmatrix}, \quad \mathbf{Z}_2 = \begin{pmatrix} 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \\ 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \mathbf{Z}_3 = \mathbf{I}_6$$

assigning the respective genetic effect i and environmental effect j to the phenotype $y_{i,j}$. We assume the variable "genotype × environment interaction" to be $(\mathbf{ge}) \sim \mathbf{N}(0, \sigma_{ge}^2\, \mathbf{GE})$ with $\mathbf{GE}$ being the covariance structure describing the similarity of genotype × environment interactions, and which needs to be specified. Let $(ge)_{i,j}$ denote the genotype × environment interaction of genotype i and environment j. As a first approach, we aim at modeling the covariance of $(ge)_{i,j}$ and $(ge)_{k,l}$ which should depend on how similar g_i and g_k are, as well as on how similar e_j and e_l are. A concept that corresponds to using products of underlying variables is to multiply the corresponding covariances:

$$\begin{aligned} \mathrm{COV}\left((ge)_{i,j}, (ge)_{k,l}\right) &= \mathrm{COV}\left(g_i, g_k\right) \cdot \mathrm{COV}\left(e_j, e_l\right) \\ &= G_{i,k} \cdot E_{j,l} \end{aligned}$$

A more detailed formal elaboration on how products of the entries of covariance matrices relate to interactions on a variable effect level can be found in the Appendix. The Kronecker product provides exactly these products of the different entries of $\mathbf{G}$ and $\mathbf{E}$, and thus a first option to define $\mathbf{GE}$ is

$$\mathbf{GE} := \mathbf{G} \otimes \mathbf{E} \tag{2}$$

It is important to highlight that Equation (2) is only valid due to the way the phenotypes are ordered. If we order first according to environment and then according to genotype, we would have to permute $\mathbf{G}$ and $\mathbf{E}$ here.

Alternatively, as a second approach, we could say that we do not focus on the covariance of the interaction between $\mathbf{g}$ and $\mathbf{e}$, but rather on the covariance of the interaction between $\mathbf{Z}_1\mathbf{g}$ and $\mathbf{Z}_2\mathbf{e}$, that is on the interaction between the two random terms of size 6×1. The variance of $\mathbf{Z}_1\mathbf{g}$ is given by $\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t$ and the variance of $\mathbf{Z}_2\mathbf{e}$ is $\mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t$, both square matrices of size 6×6. Instead of using the double index $(\mathrm{ge})_{k,l}$ for the vector $(\mathbf{ge})$ indicating genotype and environment as above, we use here only a single index indicating the position of the considered entry. This means $(\mathrm{ge})_i$ denotes the i-th entry of vector $(\mathbf{ge})$.

Again, as before, we assume that the similarity/covariance of the interaction between $(\mathbf{Z}_1\mathbf{g})_i$ and $(\mathbf{Z}_2\mathbf{e})_i$, which is $(\mathrm{ge})_i$, and the interaction between $(\mathbf{Z}_1\mathbf{g})_j$ and $(\mathbf{Z}_2\mathbf{e})_j$, which is $(\mathrm{ge})_j$, depends on the covariances $\mathrm{COV}((\mathbf{Z}_1\mathbf{g})_i, (\mathbf{Z}_1\mathbf{g})_j)$ and $\mathrm{COV}((\mathbf{Z}_2\mathbf{e})_i, (\mathbf{Z}_2\mathbf{e})_j)$ and can be modelled as the product:

$$\mathrm{COV}\left((\mathrm{ge})_i, (\mathrm{ge})_j\right) = \mathrm{COV}\left((\mathbf{Z}_1\mathbf{g})_i, (\mathbf{Z}_1\mathbf{g})_j\right) \cdot$$

$$\mathrm{COV}\left((\mathbf{Z}_2\mathbf{e})_i, (\mathbf{Z}_2\mathbf{e})_j\right) = \left(\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t\right)_{i,j} \cdot \left(\mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t\right)_{i,j}$$

Thus, we can define the covariance of their interactions as

$$\mathbf{GE} := \mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t \tag{3}$$

Equation (3) is what is most commonly used in literature addressing the topic of G × E (e.g. Jarquin et al., 2014). As we will see in the theoretical results, both operations lead to the same model, that is Equation (2) is equal to Equation (3).

3 | THEORETICAL RESULTS

We start with two theoretical results.

3.1 | First theoretical result

Let us consider a data set of n lines whose performances were measured in m environments. We include each combination of line and environment in the statistical model. Moreover, let $\mathbf{y}$ be a vector of phenotypes ordered primarily according to the genotypes and secondarily according to the environments the record was taken in. Also, let the design matrices, as described above, map the genetic value and the environment to the corresponding phenotype. Then

$$\mathbf{G} \otimes \mathbf{E} = \mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t \tag{4}$$

which means both methods give the same covariance structure $\mathbf{GE}$.

We found a similar statement in literature, in which permutation matrices are applied to the Kronecker product on the left-hand side of the equation leading to a Hadamard product on the right-hand site (e.g. Liu & Trenkler, 2008; Visick, 2000). However, we have not found Equation (4) in publications addressing the relation between Hadamard and Kronecker products.

3.1.1 | Proof of the first theoretical result

This result is a special case of the second theoretical result, which can be found below. Nevertheless, we give a separate proof for both results. In particular, the proof of the second result uses a presentation, which we introduce in the proof of the first result. Both proofs use similar arguments in which the indices are compared and entries are counted.

Let us consider the entry $(\mathbf{G} \otimes \mathbf{E})_{p,q}$. What we have to show is

$$(\mathbf{G} \otimes \mathbf{E})_{p,q} = \left(\mathbf{Z}_1\mathbf{G}\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t\right)_{p,q} \tag{5}$$

Let the design matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ be defined as before. Moreover, let the p-th entry of $(\mathbf{ge})$ be $(\mathrm{ge})_{\mathrm{i,j}}$ and let the q-th entry be $(\mathrm{ge})_{k,l}$. This means that the p-th row of $\mathbf{Z}_1$ has a 1 at position i and the q-th row of $\mathbf{Z}_1$ has a 1 at position k (the other entries of these rows are 0). Moreover, the p-th row of $\mathbf{Z}_2$ has a 1 at position j and the q-th row of $\mathbf{Z}_2$ has a 1 at position l (the other entries of these rows are 0). Thus,

$$(\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t)_{p,q} = \mathrm{G}_{i,k} \text{ and } (\mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t)_{p,q} = \mathrm{E}_{j,l}$$

and consequently

$$\left(\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t\right)_{p,q} = \mathrm{G}_{i,k} \cdot \mathrm{E}_{j,l}$$

What remains to be shown is $(\mathbf{G} \otimes \mathbf{E})_{p,q} = \mathrm{G}_{i,k} \cdot \mathrm{E}_{j,l}$. To see this, we use that

$$p = (i-1) \cdot m + j \text{ and } q = (k-1) \cdot m + l \tag{6}$$

where m is the number of environments (due to the p-th entry of $(\mathbf{ge})$ being $(\mathrm{ge})_{i,j}$ and the q-th entry being $(\mathrm{ge})_{k,l}$). Moreover, we know that

$$(\mathbf{G} \otimes \mathbf{E})_{p,q} = \mathrm{G}_{\lceil p/m \rceil, \lceil q/m \rceil} \cdot \mathrm{E}_{p \bmod m, q \bmod m} \tag{7}$$

where $\lceil p/m \rceil$ denotes the least natural number greater than or equal to the ratio p/m (rounding up; ceiling function) and where p mod m denotes the modulo operation, that is division with remainder r and here $\mathrm{r} \in \{1, \ldots, m\}$, not $\{0, \ldots, m-1\}$. Combining Equations (6) and (7) gives

$$(\mathbf{G} \otimes \mathbf{E})_{p,q} = \mathrm{G}_{i,k} \cdot \mathrm{E}_{j,l} \qquad \square$$

3.1.2 | Remark

The setup described above is based on an order of the phenotypes primarily according to line and secondarily according to environment. Of course, this can also be ordered first according to environment and then according to line. In this case, the matrices $\mathbf{Z}_1$ and $\mathbf{Z}_2$ will be adapted and the Kronecker product in Equation (4) will change to $\mathbf{E} \otimes \mathbf{G}$.

The first theoretical result describes the balanced case of having all genotype-in-environment combinations in the model and without repetitions. This corresponds exactly to the illustrating example in the Theoretical Background section. We will now go to a more general case in which no restrictions are made on the order of the entries of $\mathbf{y}$ or on the distribution of lines across environments. Moreover, we can also include repetitions. The second theoretical result is a generalization of the first result.

3.2 | Second theoretical result

Let $\mathbf{y}$ be the vector of phenotypes included in the model. An entry of $\mathbf{y}$ is given by $y_{i,j,rep}$, where i indicates the line, j the environment the record was taken in, and rep the repetition. Moreover, let $\tilde{\mathbf{Z}}_1$ be a design matrix that maps the corresponding genetic value g_i to the phenotype $y_{i,j,rep}$, let $\tilde{\mathbf{Z}}_2$ be a matrix that assigns e_j to $y_{i,j,rep}$, and let $\tilde{\mathbf{Z}}_3$ be a matrix mapping $(\mathrm{ge})_{i,j}$ to $y_{i,j,rep}$ and which is constructed as described by Bates, Mächler, Bolker, and Walker (2015) using the Khatri-Rao product

$$\tilde{\mathbf{Z}}_3 = \left(\tilde{\mathbf{Z}}_1^{\mathrm{t}} * \tilde{\mathbf{Z}}_2^{\mathrm{t}}\right)^{\mathrm{t}}.$$

Here, $(\mathbf{ge})$ includes all possible genotype-in-environment interactions and is sorted first according to genotype and second according to environment.

We consider the model

$$\mathbf{y} = \mu \mathbf{1}_{|\mathbf{y}|} + \tilde{\mathbf{Z}}_1 \mathbf{g} + \tilde{\mathbf{Z}}_2 \mathbf{e} + \tilde{\mathbf{Z}}_3 (\mathbf{ge}) + \varepsilon \tag{8}$$

with $|\mathbf{y}|$ denoting the length, that is, the number of entries of $\mathbf{y}$.

Then

$$\left(\tilde{\mathbf{Z}}_3 \left(\mathbf{G} \otimes \mathbf{E}\right) \tilde{\mathbf{Z}}_3^{\mathrm{t}}\right) = \tilde{\mathbf{Z}}_1 \mathbf{G} \tilde{\mathbf{Z}}_1^{\mathrm{t}} \circ \tilde{\mathbf{Z}}_2 \mathbf{E} \tilde{\mathbf{Z}}_2^{\mathrm{t}} \tag{9}$$

Which means we can use a Hadamard or a Kronecker formulation to specify the same model.

Before we come to the proof of the second result, please note that Equation (9) is a generalization of Equation (4).

3.2.1 | Proof of the second theoretical result

Let the v-th entry of $\mathbf{y}$ belong to a repetition of line i in environment j and let the w-th entry of $\mathbf{y}$ belong to a repetition of line k in environment l. Then the v-th row of $\tilde{\mathbf{Z}}_1$ has an entry 1 at position i. The other entries of this row are equal to 0. Moreover, the w-th row of $\tilde{\mathbf{Z}}_1$ has an entry 1 at position k. Analogously, the v-th row of $\tilde{\mathbf{Z}}_2$ has an entry 1 at position j, the w-th row of $\tilde{\mathbf{Z}}_2$ has an entry 1 at position l. Moreover, using the Khatri-Rao product definition of $\tilde{\mathbf{Z}}_3$ and the former statements on $\tilde{\mathbf{Z}}_1$ and $\tilde{\mathbf{Z}}_2$, we see that the v-th row of $\tilde{\mathbf{Z}}_3$ has a 1 at position $(i-1) \cdot m + j$, and the w-th row of $\tilde{\mathbf{Z}}_3$ has a 1 at position $(k-1) \cdot m + l$.

What we have to show is

$$\left(\tilde{\mathbf{Z}}_3 \left(\mathbf{G} \otimes \mathbf{E}\right)\ \tilde{\mathbf{Z}}_3^{\mathrm{t}}\right)_{v,w} = \left(\tilde{\mathbf{Z}}_1 \mathbf{G} \tilde{\mathbf{Z}}_1^{\mathrm{t}} \circ \tilde{\mathbf{Z}}_2 \mathbf{E} \tilde{\mathbf{Z}}_2^{\mathrm{t}}\right)_{v,w}$$

First note that according to what we described at which position the rows of the matrices have the entry 1,

$$\left(\tilde{\mathbf{Z}}_1 \mathbf{G}\, \tilde{\mathbf{Z}}_1^{\mathrm{t}}\right)_{v,w} = \mathrm{G}_{i,k} \text{ and } \left(\tilde{\mathbf{Z}}_2 \mathbf{E} \tilde{\mathbf{Z}}_2^{\mathrm{t}}\right)_{v,w} = E_{j,l}$$

Analogously $(\tilde{\mathbf{Z}}_3 (\mathbf{G} \otimes \mathbf{E})\, \tilde{\mathbf{Z}}_3^{\mathrm{t}})_{v,w}$ is the $((i-1) \cdot m + j, (k-1) \cdot m + l)$ entry of $\mathbf{G} \otimes \mathbf{E}$.
Using Equation (7) states that this is

$$(\mathbf{G} \otimes \mathbf{E})_{(i-1)\cdot m+j,(k-1)\cdot m+l} = \mathrm{G}_{\lceil((i-1)\cdot m+j)/m)\rceil, \lceil((k-1)\cdot m+l)/m\rceil} \cdot \mathrm{E}_{((i-1)\cdot m+j) \bmod m, ((k-1)\cdot m+l) \bmod m}$$

Which means

$$(\mathbf{G} \otimes \mathbf{E})_{(i-1)\cdot m+j,(k-1)\cdot m+l} = \mathrm{G}_{i,k} \cdot \mathrm{E}_{j,l} \qquad \square$$

To illustrate a case addressed by the second result, we give a simple example.

3.3 | A small example for unbalanced data

Let us consider the data

$$\mathbf{y} = \left(y_{1,1,1}, y_{1,1,2}, y_{2,2,1}, y_{3,2,1}, y_{3,3,1}, y_{2,2,2}\right)^{t}$$

consisting of phenotypes of three lines measured in three environments and some with repetitions. The entry $y_{3,2,1}$ denotes the first repetition of a record of line 3 grown in environment 2. Note that not all lines have records for each environment and not each combination is repeated. In this example, the design matrices are given by

$$\tilde{\mathbf{Z}}_1 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}, \tilde{\mathbf{Z}}_2 = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix},$$

$$\tilde{\mathbf{Z}}_3 = \begin{pmatrix} 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \end{pmatrix}$$

Use random matrices **G** and **E** to see that Equation (9) is satisfied. Note that $\tilde{\mathbf{Z}}_3 = (\tilde{\mathbf{Z}}_1^t * \tilde{\mathbf{Z}}_2^t)^t$ and recall that the Khatri-Rao product is not commutative, that is $(\tilde{\mathbf{Z}}_1^t * \tilde{\mathbf{Z}}_2^t)^t \neq (\tilde{\mathbf{Z}}_2^t * \tilde{\mathbf{Z}}_1^t)^t$. The order of the multiplication has to be aligned with the order of the Kronecker product on the left-hand side of Equation (9). If we would use $\mathbf{E} \otimes \mathbf{G}$ instead of $\mathbf{G} \otimes \mathbf{E}$, the third design matrix would be given by $\tilde{\mathbf{Z}}_3 = (\tilde{\mathbf{Z}}_2^t * \tilde{\mathbf{Z}}_1^t)^t$.

We will now consider a wheat and a maize data set to show that the inclusion of available data from other environments may be beneficial, also in the case that we are only interested in predicting the performance of lines grown in a certain environment.

4 | MATERIALS AND METHODS

Additionally to the theoretical results, we present some empirical results on predictive ability of models including G × E interaction. In the following, we will explain the data structure and the cross validation scenarios which we used.

TABLE 1 Phenotypic correlation of the yield records of the 599 lines across the four environments (wheat data)

	E1	E2	E3	E4
E1	1	−0.02	−0.19	−0.12
E2		1	0.66	0.41
E3			1	0.39
E4				1

TABLE 2 Phenotypic correlation of the yield records of the 722 lines planted in all four environments (maize data)

	E1	E2	E3	E4
E1	1	0.54	0.33	0.42
E2		1	0.37	0.47
E3			1	0.49
E4				1

4.1 | Data

4.1.1 | Wheat data

We used the wheat data set published by Crossa et al. (2010) which offers genomic data in the form of 1279 presence/absence markers of 599 wheat lines. The phenotypic data provides yield records of these 599 lines grown under four different environmental conditions. The correlation of the phenotypes of the 599 lines grown in the four environments is presented in Table 1. For more information on the data see Crossa et al. (2010).

4.1.2 | Maize data

We additionally considered a maize data set which has been published by Sousa et al. (2017). The data set USP provides phenotypic records of different traits of 739 lines measured in four environments. We restrict our considerations to yield. Here, 722 lines had records from all four environments, and 17 lines were not observed at each location. The genomic relationship matrix GB was used as provided. The phenotypic correlations of the lines overlapping between each pair of environment are presented in Table 2. For more information on the data, please see Sousa et al. (2017).

4.2 | Cross validations, statistical models and covariance matrices

To illustrate the properties of G × E models, we considered three different types of cross validations. All predictions

TABLE 3 Predictive abilities for the wheat data. Mean Pearson correlation of the predicted values and the phenotypes of the test set (60 lines) for the different cross validation scenarios. SC1 is within environment (E1, E2, E3 or E4), ignoring the data from the other environments. SC2 and SC3 use a G × E model and the data of the other environments. The difference between both is whether the phenotypic data of the 60 lines of the test set under different environmental conditions is included (SC2) or not (SC3). Moreover, the E and I indicate whether the identity matrix (I) is used as environmental covariance or whether the environmental covariance is estimated from the phenotypic covariance across environments in the training data (E)

Environment	SC1	SC2-I	SC2-E	SC3-I	SC3-E
E1	0.517 ± 0.013	0.432 ± 0.016	0.542 ± 0.012	0.461 ± 0.017	0.516 ± 0.013
E2	0.512 ± 0.015	0.562 ± 0.015	0.657 ± 0.010	0.491 ± 0.015	0.532 ± 0.015
E3	0.382 ± 0.014	0.448 ± 0.016	0.528 ± 0.014	0.385 ± 0.017	0.391 ± 0.016
E4	0.468 ± 0.017	0.493 ± 0.012	0.541 ± 0.011	0.453 ± 0.014	0.452 ± 0.013

were performed using the BGLR package (Perez & de los Campos, 2014).

For the wheat data set, the genomic relationship matrix was constructed as $\mathbf{G} := \mathbf{M}\mathbf{M}^t/p$ (VanRaden, 2008) with $\mathbf{M}$ the 599 × 1279 marker matrix, and p the number of markers (here 1279). For the maize data set, the GBLUP matrix was used as provided by Sousa et al. (2017).

Moreover, when $\mathbf{E}$ was included in the model, we considered two cases. The first case was $\mathbf{E} := \mathbf{I}$ the identity matrix. For the second case, the lines of the training set were used to estimate the pairwise covariances of the phenotypes across different environments and the R function nearPD of the package "Matrix" (Bates and Maechler, 2019) was applied to guarantee the positive-semi-definiteness of $\mathbf{E}$.

4.2.1 | Scenario SC1: Standard within environment CV

The first cross validation was within environment. For the wheat data, a test set of 60 lines was randomly drawn and the genetic values were predicted using the remaining 539 lines. This was repeated 50 times. The average Pearson correlation of predictions and phenotypes of the test set is reported in Table 3. The statistical model used was

$$\mathbf{y} = \mu\mathbf{1}_{599} + \tilde{\mathbf{Z}}_1\mathbf{g} + \varepsilon$$

with $\mathbf{g} \sim \mathbf{N}(0, \sigma_g^2\,\mathbf{G})$ and $\varepsilon \sim \mathbf{N}(0, \sigma_\varepsilon^2\,\mathbf{I})$. The Pearson correlation measured was $\text{cor}(\mathbf{y}, \tilde{\mathbf{Z}}_1\hat{\mathbf{g}})$ on the test set, that is the set of 60 lines whose genetic values were predicted from the phenotypes of the remaining 539 lines.

The maize data set was used analogously, but with a test set of 73 lines.

4.2.2 | Scenario SC2: Using data from other environments including the records of the same lines

For the second cross validation, we included the record of the lines from other environment, but the test set is still coming from one environment. For each of the 50 repetitions, 60 lines were drawn as test set from environment i and predicted by the data of the remaining 539 lines in environment i and the data of all 599 lines in the other three environments. The statistical model used was

$$\mathbf{y} = \mu\mathbf{1}_{599} + \tilde{\mathbf{Z}}_1\mathbf{g} + \tilde{\mathbf{Z}}_2\mathbf{e} + \tilde{\mathbf{Z}}_3(\mathbf{ge}) + \varepsilon$$

with $\mathbf{g}$ and ε as before and $\mathbf{e} \sim \mathbf{N}(0, \sigma_e^2\,\mathbf{E})$, $(\mathbf{ge}) \sim \mathbf{N}(0, \sigma_{ge}^2\,\mathbf{G} \otimes \mathbf{E})$. The design matrices were generated according to the data structure as earlier described. The Pearson correlation measured was $\text{cor}(\mathbf{y},\ \tilde{\mathbf{Z}}_1\hat{\mathbf{g}} + \tilde{\mathbf{Z}}_3(\widehat{\mathbf{ge}}))$ for the test set. Since the test set was selected from the same environment, the correlation does not depend on whether the environmental effects are included. Two variants were calculated. The first variant was $\mathbf{E} := \mathbf{I}$. In the second variant, $\mathbf{E}$ was estimated by the phenotypic correlation across environments of the training set (as described above).

The procedure for the maize data set was analogous, but with a test set of 73 lines.

4.2.3 | Scenario SC3: Using data from other environments without the records of the same lines

Scenario SC3 was analogous to SC2, but without using any record of lines in the test set. This means that having drawn a test set in an environment, we used the records of the remaining 539 lines from all four environments as training data.

The procedure for the maize data set was analogous, but with a test set of 73 lines.

5 | EMPIRICAL RESULTS

5.1 | Wheat data

The first cross validation scenario SC1 was based on a within environment prediction. In the second scenario SC2, we were again interested in the predictive ability of yield in a particular environment. However, instead of using only the 539 records of the training set of the particular environment, we additionally used the records of the 599 lines in the other three environments. We used a model including G × E effects in two variations. In the first version SC2-I, we used the identity matrix $\mathbf{E} := \mathbf{I}$. In the second version SC2-E, we estimated the pairwise phenotypic covariances between environments from the training set. The third cross validation SC3, was the same as SC2, but we did not include the phenotypic data of the lines of the test set grown in the other environments. Instead, we only used the data of the remaining 539 lines in the four environments as training set. The distinction between SC2 and SC3 allows separating the effect of having the same lines under different environmental conditions from the general effect of including G × E and more environments in the training set. We also used the variants SC3-I and SC3-E.

The results are summarized in Table 3. Comparing SC1 to SC2-I, we see that the inclusion of the other environments and modeling G × E increased the predictive ability in environments 2, 3, and 4. However, the predictive ability in environment 1 was drastically reduced (from 0.517 to 0.432). This may be a result of the negative phenotypic correlation, which the phenotypes of environment 1 have with the phenotypes of the other environments (see Table 1, Materials and Methods) and not having included this aspect in the environmental covariance.

Moreover, when we compare SC1 to SC3-I, we see that the increase in predictive ability given by SC2-I is lost when the yield performance of the test set lines under different environmental conditions is not included in the training set (compare SC2-I to SC3-I). Thus, the main increase of SC2-I compared to SC1 seems to be a result of having some records of the performance of the test set in the training set. Also the negative impact for the prediction of a test set of environment 1 is reduced when the records of test set lines in other environments are not included in the training set (compare E1, 0.432 for SC2-I to 0.461 for SC3-I).

Considering the cross validations SC2-E and SC3-E, we again see – as for SC2-I and SC3-I and E2, E3, E4 – that SC2-E leads to higher predictive abilities than SC3-E. Moreover, when comparing both to their counterpart with the identity matrix, we see that estimating $\mathbf{E}$ from the training set is beneficial and also improves the prediction of lines in environment 1 (from 0.517 in SC1 to 0.542 in SC2-E). Comparing SC1 to SC3-E, the difference in predictive ability is rather small indicating again that a main driver to increase predictive ability is including records of the test set from other environments in the training set.

An important point to highlight is that the predictive ability for SC2-E was in environment 2 and 3 below the maximal correlation of the phenotypes across environments, which was here 0.66 for environments 2 and 3 (see Table 1). However, for environment 1 and 4, the predictive ability using SC2-E was above the maximal (absolute) phenotypic correlation with these environments. For instance, E2 has the highest phenotypic correlation with the data of E4 (0.41). The predictive ability for E4 increased from 0.468 (SC1) to 0.541 (SC2-E). For environments E1 and E4, SC2-E improved the prediction compared to simply using the phenotypic records from the environment that has the highest correlation with the test environment to predict the test set.

In this context of using the data of the other environments including records of the test set lines, see also the results of Martini et al. (2016). The authors used one environment to identify relevant interactions of an epistasis model and then subnetworks of selected interactions to define a relationship matrix. The latter was subsequently used to predict within another environment. The predictive abilities obtained were very similar to those of SC2-E, which used the same data basis.

5.2 | Maize data

The predictive abilities for the maize data set are summarized in Table 4. The patterns observed are similar to those seen on the wheat data set in the sense that SC2-E is the best scenario for the prediction of three of four environments. An important difference to the results for the wheat data set, is that in none of the cases a predictive ability above the maximal phenotypic correlation with one of the other environments was reached (Table 2). Moreover, in environment 4, the predictive ability of scenario SC3 were a little bit higher than that those of the corresponding SC2 scenarios.

6 | DISCUSSION

We have shown that the covariance structures of interactions expressed by the Hadamard products

TABLE 4 Predictive abilities for the maize data. Mean Pearson correlation of the predicted values and the phenotypes of the test set (73 lines) for the different cross validation scenarios. SC1 is within environment (E1, E2, E3 or E4), ignoring the data from the other environments. SC2 and SC3 use a G × E model and the data of the other environments. The difference between both is whether the phenotypic data of the 73 lines of the test set under different environmental conditions is included (SC2) or not (SC3). Moreover, the E and I indicate whether the identity matrix (I) is used as environmental covariance or whether the environmental covariance is estimated from the phenotypic covariance across environments in the training data (E)

Environment	SC1	SC2-I	SC2-E	SC3-I	SC3-E
E1	0.278 ± 0.0122	0.313 ± 0.017	0.323 ± 0.017	0.308 ± 0.017	0.308 ± 0.016
E2	0.334 ± 0.014	0.347± 0.017	0.350 ± 0.017	0.342 ± 0.018	0.344 ± 0.018
E3	0.322 ± 0.014	0.367 ± 0.011	0.377 ± 0.011	0.307 ± 0.012	0.307 ± 0.012
E4	0.416 ± 0.014	0.452 ± 0.010	0.455 ± 0.010	0.466 ± 0.010	0.466 ± 0.011

$\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t$ can equally be written as the Kronecker product $\mathbf{G} \otimes \mathbf{E}$ (Equation (4) and with the design matrices $\mathbf{Z}_1,\ \mathbf{Z}_2$ as described). An important aspect in this context is the order of the phenotypes. For Equation (4), we used an order of primarily according to the genotypes and secondarily according to the environments. Ordering for instance first according to environments and then according to genotypes would lead to an analogous equation, but with $\mathbf{G} \otimes \mathbf{E}$ commuted to $\mathbf{E} \otimes \mathbf{G}$ in Equation (4). For unbalanced data, when including repetitions, or for an arbitrary order of the phenotypes, the design matrices have to be adapted, but an analogous, more general identity holds (Equation (9)). Thus, if the design matrices are defined according to the order of the phenotypes, both approaches give exactly the same covariance structure.

Having the equivalence of both formulations in mind, it can be interpreted in parts as coincidence that the Hadamard formulation can be found more often in literature addressing G × E, but that the Kronecker formulation has been used for modeling specific combining ability.

6.1 | Potential computational advantages - Kronecker and Hadamard products and their inverses

Each of the two different presentations may have computational (dis)advantages compared to the other and depending on which mathematical operations should be performed. A classical numerical problem related to the mixed model equations is inverting the covariance matrix. Here, we see that that the Kronecker product may give some advantages since it obeys the identity

$$(\mathbf{G} \otimes \mathbf{E})^{-1} = \mathbf{G}^{-1} \otimes \mathbf{E}^{-1}$$

if all matrices are invertible (e.g. Neudecker, 1969). Since the running time of an inversion of a square matrix of size $n \times n$ grows at least as a polynomial of degree 2 (e.g. Petković & Stanimirović, 2009; Wilf, 2002), the right-hand side is computationally simpler due to the lower dimensionality of the matrices to be inverted. Considering the formulation $(\mathbf{Z}_1\mathbf{G}\,\mathbf{Z}_1^t \circ \mathbf{Z}_2\mathbf{E}\mathbf{Z}_2^t)^{-1}$, a similar equality is not given.

6.2 | The issue of variable coding when modeling interactions by products of predictor variables

We illustrated the equivalence of multiplying entries of the genomic relationship and the environmental covariance, and an approach of using explicitly products of markers and environmental covariables as additional predictors with their own interaction effect (see Appendix). An analogous result has earlier been shown in the context of the interactions within a class of variables, more concrete locus × locus interactions (Jiang & Reif, 2015; Martini et al., 2016). When using products of the independent variables as additional predictors to model interactions between them, the variable coding may have an impact on the prediction outcome when penalized regressions are used (Martini et al., 2017; Martini et al., 2019). Different codings have been discussed, for instance approaching orthogonal estimates of variances (Vitezica et al., 2017). Similar to the locus × locus interaction, the presented equivalences are true for any variable coding (Martini et al., 2016), but the fact that the variable coding will have an impact on the prediction when using a penalized regression method such as RRBLUP/GBLUP equally applies.

6.3 | Practical implications for sparse testing designs

The results illustrate once more that including G × E effects into prediction models can increase predictive ability (compare to e.g. Burgueño et al., 2012;

Cuevas et al., 2016; Jarquin et al., 2014). Concerning the practical application of G × E models, we showed that for the considered methods, the main drivers for the increase in predictive ability are (i) specifying the environmental covariances correctly **and** (ii) having the same lines grown in the other environments (The latter has been highlighted for instance by Burgueño et al., 2012). This becomes evident when considering Tables 3 and 4, for which the scenario SC2-E had the highest predictive ability for seven out of the eight considered cases. For sparse testing approaches in which the set of lines is partitioned across environments, this means that all lines should be observed in at least one environment but also that the environmental covariance **E** should be determined. If no data on environmental variables is available but there is sufficient overlap of lines between environments, **E** can be estimated as phenotypic (or estimated genetic) covariance.

6.4 | Outlook

The G × E models have been demonstrated to improve predictive ability in many instances. One important application of these approaches is to define "optimal" sparse testing designs which allow reducing the testing effort without a big loss in accuracy. An investigation of different sparse testing designs together with a comparison of different G × E models may be of interest for the plant breeding community.

ACKNOWLEDGMENTS

We are thankful for the financial support provided by the Bill & the Melinda Gates Foundation (for maize and wheat breeding programs), the CIMMYT CGIAR CRP (maize and wheat), as well the USAID projects. 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 would like to thank two unknown reviewers for their valuable comments, especially for pointing at the possible advantage of the Kronecker formulation for inverting the matrices. Moreover, we thank Francisco Rosales for comments on the manuscript.

ORCID

Johannes W. R. Martini https://orcid.org/0000-0003-0628-6794
Jose Crossa https://orcid.org/0000-0001-9429-5855
Fernando H. Toledo https://orcid.org/0000-0003-0158-643X

REFERENCES

Acosta-Pech, R., Crossa, J., de los Campos, G., Teyssèdre, S., Claustres, B., Pérez-Elizalde, S., & Pérez-Rodríguez, P. (2017). Genomic models with genotype × environment interaction for predicting hybrid performance: An application in maize hybrids. *Theoretical and Applied Genetics*, *130*(7), 1431–1440.

Basnet, B. R., Crossa, J., Dreisigacker, S., Pérez-Rodríguez, P., Manes, Y., Singh, Ravi P., … Murua, M. (2019). Hybrid wheat prediction using genomic, pedigree, and environmental covariables interaction models. *The Plant Genome*, *12*(1).

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*(1).

Bates, D., & Mächler, M. (2019). Matrix: Sparse and dense matrix classes and methods. *R package version*, *1*, 2–17. Retrieved from https://CRAN.R-project.org/package=Matrix.

Burgueño, J., Crossa, J., Cornelius, P. L., Trethowan, R., McLaren, G., & Krishnamachari, A. (2007). Modeling additive × environment and additive × additive × environment using genetic covariances of relatives of wheat genotypes. *Crop Science*, *43*, 311–320.

Burgueño, J., de los Campos, G., Weigel, K., & Crossa, J. (2012). Genomic prediction of breeding values when modeling genotype × environment interaction using pedigree and dense molecular markers. *Crop Science*, *52*(2), 707–719.

Cornelius, P. L., Crossa, J., & Seyedsadr, M. S. (1996). Statistical tests and estimators of multiplicative models for genotype-by-environment interaction. In M. S. Kang & H. G. Gauch (Eds.), *Genotype-by-environment interaction* (pp. 199–234). Boca Raton: CRC Press.

Cornelius, P. L., & Crossa, J. (1999). Prediction assessment of shrinkage estimators of multiplicative models for multi-environment cultivar trials. *Crop Science*, *39*, 998–1009.

Cornelius, P. L., Crossa, J., Seyedsadr, M. S., Liu, G., & Viele, K. (2001). Contributions to multiplicative model analysis of genotype-environment data. Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, Proceedings, American Statistical Association, 2001.

Crossa, J., & Cornelius, P. L. (1997). Site regression and shifted multiplicative model clustering of cultivar trials sites under heterogeneity of error variances. *Crop Science*, *37*, 406–415.

Crossa, J., van Eeuwijk, F., Vargas, M., & Cornelius, P. L. (2001). Linear, bilinear and linear bilinear models for analyzing genotype-environment interaction. Statistical Consulting Section, American Statistical Association, Joint Statistical Meetings, Proceedings, American Statistical Association, 2001.

Crossa, J., de los Campos, G., Pérez, P., Gianola, D., Burgueño, J., Araus, J. L., … Braun, H.- J. (2010). Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. *Genetics*, *18*(2), 713–724.

Crossa, J., Pérez, P., de los Campos, G., Mahuku, G., Dreisigacker, S., & Magorokosho, C. (2011). Genomic selection and prediction in plant breeding. *Journal of Crop Improvement*, *25*(3), 239–261.

Crossa, J., Pérez-Rodríguez, P., Cuevas, J., Montesinos-López, O., Jarquín, D., de los Campos, G., … Varshney, R. K. (2017). Genomic selection in plant breeding: Methods, models, and perspectives. *Trends in Plant Science*, *22*(11), 961–975.

Crossa, J., Burgueno, J., Cornelius, P. L., McLaren, G., Trethowan, R., & Krishnamachari, A. (2006). Modeling genotype × environment interaction using additive genetic covariances of relatives for predicting breeding values of wheat genotypes. *Crop Science*, *46*, 1722–1733.

Cuevas, J., Crossa, J., Soberanis, V., Pérez-Elizalde, S., Pérez-Rodríguez, P., de los Campos, G., … Burgueño, J. (2016). Genomic prediction of genotype × environment interaction kernel regression models. *The Plant Genome*, *9*(3).

Cuevas, J., Crossa, J., Montesinos-López, O. A., Burgueño, J., Pérez-Rodríguez, P., & de los Campos, G. (2017). Bayesian genomic prediction with genotype× environment interaction kernel models. *G3: Genes, Genomes, Genetics*, *7*(1), 41–53.

de los Campos, Gustavo, Gianola, Daniel, Rosa, Guilherme J. M., Weigel, Kent A., & Crossa, José (2010). Semi-parametric genomic-enabled prediction of genetic values using reproducing kernel Hilbert spaces methods. *Genetics Research*, *92*(4), 295–308.

de los Campos, G., Gianola, D., & Rosa, G. J. M. (2009). Reproducing kernel Hilbert spaces regression: A general framework for genetic evaluation. *Journal of Animal Science*, *87*(6), 1883–1887.

Gao, N., Martini, J. W. R., Zhang, Z., Yuan, X., Zhang, H., Simianer, H., & Li, J. (2017). Incorporating gene annotation into genomic prediction of complex phenotypes. *Genetics*, *207*(2), 489–501.

Gonzales-Barrios, P., Diaz-Garcia, L., & Gutierrez, L. (2019). Mega-environmental design: Using genotype × environment interaction to optimize resources for cultivar testing. *Crop Science*, *59*(5), 1899–1915.

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

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

Jannink, J.-L., Lorenz, A. J., & Iwata, H. (2010). Genomic selection in plant breeding: From theory to practice. *Briefings in Functional Genomics*, *9*(2), 166–177.

Jarquin, D., Crossa, J., Lacaze, X., Du Cheyron, P., Daucourt, J., Lorgeou, J., … de los Campos, G. (2014). A reaction norm model for genomic selection using high-dimensional genomic and environmental data. *Theoretical and Applied Genetics*, *127*(3), 595–607.

Jiang, Y., & Reif, J. C. (2015). Modeling epistasis in genomic selection. *Genetics*, *201*(2), 759–768.

Kolda, T. G., & Bader, B. W. (2009). Tensor decompositions and applications. *SIAM review*, *51*(3), 455–500.

Liu, S., & Trenkler, G. (2008). Hadamard, Khatri-Rao, Kronecker and other matrix products. *International Journal of Information and Systems Sciences*, *4*(1), 160–177.

Lopez-Cruz, M., Crossa, J., Bonnett, D., Dreisigacker, S., Poland, J., Jannink, J.-L., … de los Campos, G. (2015). Increased prediction accuracy in wheat breeding trials using a marker × environment interaction genomic selection model. *G3: Genes, Genomes, Genetics*, *5*(4), 569–582.

Martini, J. W. R., Wimmer, V., Erbe, M., & Simianer, H. (2016). Epistasis and covariance: How gene interaction translates into genomic relationship. *Theoretical and Applied Genetics*, *129*(5), 963–976.

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*(1), 3.

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, Genetics*, *9*(4), 1117–1129.

Martini, J. W. R., Toledo, F. H., & Crossa, J. (2020). On the approximation of interaction effect models by Hadamard powers of the additive genomic relationship. *Theoretical Population Biology*, *132*, 16–23.

Meuwissen, T. H. E., Hayes, B. J., & Goddard, M. E. (2001). Prediction of total genetic value using genome-wide dense marker maps. *Genetics*, *157*(4), 1819–1829.

Neudecker, H. (1969). A note on Kronecker matrix products and matrix equation systems. *SIAM Journal on Applied Mathematics*, *17*(3), 603–606.

Ober, U., Erbe, M., Long, N., Porcu, E., Schlather, M., & Simianer, H. (2011). Predicting genetic values: A kernel-based best linear unbiased prediction with genomic data. *Genetics*, *188*(3), 695–708.

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

Perez-Rodriguez, P., Crossa, J., Bondalapati, K., De Meyer, G., Pita, F., & de los Campos, G. (2015). A pedigree-based reaction norm model for prediction of cotton yield in multienvironment trials. *Crop Science*, *55*(3), 1143–1151.

Perez-Rodriguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., … Dreisigacker, S. (2017). Single-step genomic and pedigree genotype × environment interaction models for predicting wheat lines in international environments. *The Plant Genome*, *10*(2).

Petković, M. D., & Stanimirović, P. S. (2009). Generalized matrix inversion is not harder than matrix multiplication. *Journal of computational and applied mathematics*, *230*(1), 270–282.

R Core Team. (2013). *R: A language and environment for statistical computing*. Vienna: R Core Development Team.

Sousa, M. B., Cuevas, J., de Oliveira Couto, E. G., Pérez-Rodríguez, P., Jarquín, D., Fritsche-Neto, R., … Crossa, J. (2017). Genomic-enabled prediction in maize using kernel models with genotype × environment interaction. *G3: Genes, Genomes, Genetics*, *7*(6), 1995–2014.

Sukumaran, S. et al. (2017). Genomic prediction with pedigree and genotype × environment interaction in spring wheat grown in South and West Asia, North Africa, and Mexico. *G3: Genes, Genomes, Genetics*, *7*(2), 481–495.

VanRaden, P. M. (2008). Efficient methods to compute genomic predictions. *Journal of Dairy Science*, *91*(11), 4414–4423.

Varona, L., Legarra, A., Toro, M. A., & Vitezica, Z. G. (2018). Non-additive effects in genomic selection. *Frontiers in Genetics*, *9*, 78.

Visick, G. (2000). A quantitative version of the observation that the Hadamard product is a principal submatrix of the Kronecker product. *Linear Algebra and Its Applications*, *304*(1-3), 45–68.

Vitezica, Z. G., Legarra, A., Toro, M. A., & Varona, L. (2017). Orthogonal estimates of variances for additive, dominance, and epistatic effects in populations. *Genetics*, *206*(3), 1297–1307.

Wilf, H. S. (2002). *Algorithms and complexity*. Boca Raton: CRC Press.

Yao, C., de los Campos, G., VandeHaar, M. J., Spurlock, D. M., Armentano, L. E., Coffey, M., … Weigel, K. A. (2017). Use of genotype × environment interaction model to accommodate genetic heterogeneity for residual feed intake, dry matter intake, net energy in milk, and metabolic body weight in dairy cattle. *Journal of Dairy Science*, *100*(3), 2007–2016.

How to cite this article: Martini JWR, Crossa J, Toledo FH, Cuevas J. On Hadamard and Kronecker products in covariance structures for genotype × environment interaction. *Plant Genome*. 2020;13:e20033. https://doi.org/10.1002/tpg2.20033

APPENDIX

Products of entries of covariance matrices and explicit interactions

We stated in the main text that we use a model in which

$$\mathrm{COV}\ \left(ge_{ij}, ge_{kl}\right) = \mathrm{COV}\left(g_i, g_k\right) \cdot \mathrm{COV}\ \left(e_j, e_l\right) = G_{i,k} \cdot E_{j,l}$$

First note that given **A** and **B** are both positive semidefinite, $\mathbf{A} \otimes \mathbf{B}$ and $\mathbf{A} \circ \mathbf{B}$ are both positive semidefinite (if the Hadamard product is defined), which means that these matrix operations give valid covariance structures. A question that may come up is why the product of the entries is a reasonable covariance structure. The matrices **G** and **E** are usually derived from variables such as an $n \times t$ marker data **M** or an $m \times s$ matrix **N** giving environmental variables. A common definition of the covariance structures is then for instance $\mathbf{G} := \mathbf{M}\mathbf{M}^t$ and $\mathbf{E} := \mathbf{N}\mathbf{N}^t$. As before, n is the number of lines and m is the number of environments. Moreover, t and s represent the number of variables considered to characterize **G** or **E**, respectively. We can interpret **g** and **e** of Equation (1) as being a result of

$$\mathbf{g} := \mathbf{M}\boldsymbol{\beta}_1 \text{ and } \mathbf{e} := \mathbf{N}\boldsymbol{\beta}_2$$

with $\boldsymbol{\beta}_1$, $\boldsymbol{\beta}_2$ independent and $\boldsymbol{\beta}_1 \sim \mathbf{N}(0, \sigma^2_{\beta_1}\ \mathbf{I})$, $\boldsymbol{\beta}_2 \sim \mathbf{N}(0, \sigma^2_{\beta_2}\ \mathbf{I})$. Moreover, let us model the interaction of two variables by adding an effect for the product of the two variables

$$y_{i,j} = \mu + \mathbf{M}_{i,\bullet}\ \boldsymbol{\beta}_1 + \mathbf{N}_{j,\bullet}\ \boldsymbol{\beta}_2 + \sum_{r=1,\ldots,t}\ \sum_{u=1,\ldots,s} M_{i,r} \cdot N_{j,u}\ \cdot \beta_{3,r,u} + \varepsilon_{i,j}$$

The term $\sum_{r=1,\ldots,t;u=1,\ldots,s} M_{i,r} \cdot N_{j,u} \cdot \beta_{3,r,u}$ with $\boldsymbol{\beta}_3 \sim \mathbf{N}(0, \sigma^2_{\beta_3}\ \mathbf{I})$ corresponds then to $(ge)_{i,j}$. We investigate what the induced covariance structure for **(ge)** looks like:

$\mathrm{COV}\ (ge_{i,j}, ge_{k,l}) = \mathrm{COV}(\sum_{r=1,\ldots,t;u=1,\ldots,s} M_{i,r} \cdot N_{j,u} \cdot \beta_{3,r,u}, \sum_{r=1,\ldots,t;u=1,\ldots,s} M_{k,r} \cdot N_{l,u}\ \cdot \beta_{3,r,u})$. Due to the linearity of the COV and the independence of $\beta_{3,r,u}$, this can be written as

$$\mathrm{COV}\ \left(ge_{ij}, ge_{kl}\right) = \sigma^2_{\beta_3} \sum_{r=1,\ldots,t}\ \sum_{u=1,\ldots,s} M_{i,r} \cdot N_{j,u} \cdot M_{k,r} \cdot N_{l,u} = \sigma^2_{\beta_3} \left(\sum_{r=1,\ldots,t} M_{i,r} \cdot M_{k,r} \right) \cdot \left(\sum_{u=1,\ldots,s} N_{j,u} \cdot N_{l,u} \right) = \sigma^2_{\beta_3} G_{i,k} \cdot E_{j,l}$$

This illustrates that the product structure of multiplying the covariance matrices element-wise is equivalent to modeling interactions by products of the underlying variables used to define **G** and **E**. Including products of predictor variables as additional predictors is a common way to model interactions. However, there are other possibilities, for instance modeling an independent effect for each possible category of each pair of markers (e.g. Martini et al., 2017) or other “non-additive” relationship models such as the Gaussian kernel (e.g. Crossa et al., 2010; de los Campos et al., 2009; Ober et al., 2011).