RESEARCH ARTICLE

Lost in translation: On the impact of data coding on penalized regression with interactions

Johannes WR Martini[1*], Francisco Rosales[2], Ngoc-Thuy Ha[3], Thomas Kneib[4], Johannes Heise[5] and Valentin Wimmer[1]

*Correspondence:
jmartin2@gwdg.de
[1]KWS SAAT SE, Einbeck, Germany
Full list of author information is available at the end of the article

Abstract

Background Penalized regression approaches are standard tools in quantitative genetics. It is known that the fit of an *ordinary least squares* (OLS) regression is independent of certain transformations of the coding of the predictor variables, and that the standard mixed model *ridge regression best linear unbiased prediction* (RRBLUP) is neither affected by translations of the variable coding, nor by global scaling. However, it has been reported that an extended version of this mixed model, which incorporates interactions by products of markers as additional predictor variables, indeed is affected by translations of the marker coding.

Results In this work, we identify the cause of this loss of invariance in a general context of penalized regression on polynomials in the predictor variables. We show that in most cases, translating the coding of the predictor variables has an impact on effect estimates, with the exception of the situation in which only the size of the coefficients of monomials of highest degree are penalized. The invariance of RRBLUP can thus be considered as a special case of this setting, with a polynomial of degree 1, where the size of the fixed effect (degree 0) is not penalized but all coefficients of monomials of degree 1 are. The extended RRBLUP which includes interactions is not invariant to translations, since it does not only penalize interactions (degree 2), but also additive effects (degree 1). Our observation are valid for general penalized regression settings, for instance, the ℓ_1 penalized LASSO method.

Conclusion Our results give a general insight into the behavior of penalized regressions. The fact that coding translations alter the estimates of interaction effects, provides an additional reason for interpreting the biological meaning of estimated interaction effects with caution. Moreover, this problem does not only apply to gene by gene interactions, but also to other types of interactions whose covariance is modeled with Hadamard products of covariance matrices (for instance gene by environment interactions).

Keywords: epistasis; extended GBLUP; coding-dependence?

Background

Genomic prediction is the prediction of properties of individuals from their genetic data. It is a crucial ingredient of modern breeding programs [1–5]. The traditional quantitative genetics theory is built upon linear models in which allele effects are usually modeled additively [6]. In more detail, the standard model to represent the effect of the genotype on the phenotype is given by

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

where $\mathbf{y}$ is the $n \times 1$ vector of the phenotypic observations of n individuals and $\mathbf{1}_n$ an $n \times 1$ vector with each entry equal to 1. Moreover, μ is the y-intercept, and $\mathbf{M}$ the $n \times p$ matrix describing the marker states of n individuals at p loci. Dealing with single nucleotide polymorphisms (SNPs) and a diploid species, the entries $M_{i,j}$ can for instance be coded as 0 (**aa**), 1 (**aA** or **Aa**) or 2 (**AA**) counting the occurrence of the reference allele **A**. The $p \times 1$ vector $\boldsymbol{\beta}$ represents the allele substitution effects of the p loci, and $\boldsymbol{\epsilon}$ the $n \times 1$ error vector. For single marker regression, which may for instance be used in genome-wide association studies (GWAS), we can apply ordinary least squares regression to determine the estimated (predicted) $\hat{\beta}$. However, in approaches of genomic prediction, we model the effects of many different loci simultaneously and the number of markers p is usually much larger than the number of observations n. To reduce overfitting and to deal with a large number of predictor variables, different methods have been applied in the last decades, among which *ridge regression best linear unbiased prediction* (RRBLUP) is the most popular [7]. RRBLUP penalizes the squared ℓ_2 norm of $\boldsymbol{\beta}$ and is built on the additional model specifications of μ being a fixed unknown parameter, $\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\beta \mathbf{I}_p)$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2_\epsilon \mathbf{I}_n)$ with $\mathbf{I}_n$ the identity matrix of dimension n. With an approach of maximizing a certain density, these assumptions allow to derive the optimal penalty factor as the ratio of the variance components $\lambda := \frac{\sigma^2_\epsilon}{\sigma^2_\beta}$ [8–10]. Please note here again that the fixed effect μ is not penalized in RRBLUP, which means that this method is not a pure ridge regression but actually a mixed model in which the size of μ is not penalized, but the entries of $\boldsymbol{\beta}$ are. This mixed model RRBLUP is also called *genomic best linear unbiased prediction* (GBLUP) when it is reformulated with $\mathbf{g} := \mathbf{M}\boldsymbol{\beta}$, and thus $\mathbf{g} \sim \mathcal{N}(0, \sigma^2_\beta \mathbf{M}'\mathbf{M})$.

It is known that translations of the marker coding, that is, subtracting a constant m_i from the i-th column of $\mathbf{M}$, does not change the predictions $\hat{\mathbf{y}}$ of an OLS regression (provided it is well-defined). This invariance also holds for RRBLUP, when the penalty factor remains fixed. Also when modeling interactions by products of two predictor variables, that is when fitting the coefficients of a polynomial of degree 2 to the data, OLS predictions are not affected by translations of the marker coding. Contrarily, the predictions of its penalized regression analogue *extended genomic best linear unbiased prediction* (eGBLUP) are sensitive to a translation of the coding [11, 12].

In this work, we address the question of why the penalized regression method is affected by translations of the marker coding when a polynomial function of higher degree is used. We start with a short summary of the different methods.

Theory: Specification of regression methods

If an expression includes an inverse of a matrix, we implicitly assume that the matrix is invertible for the respective statement, also if not mentioned explicitly. Analogously, some statements for OLS may implicitly assume that a unique estimate exists, which, for instance, implicitly restricts to cases of $n > p$ for OLS. Moreover, please note that we use the term "estimated" for all quantities, independent of whether it has been assumed to be fixed or random in the model.

Additive effect regression

The additive effect model has already been presented in Eq. (1).

OLS The ordinary least squares approach determines $\hat{\boldsymbol{\beta}}$ by minimizing the squared residuals:

$$\begin{pmatrix}\hat{\mu}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}_{\text{OLS}} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n}(y_i - \mathbf{M}_{i,\bullet}\boldsymbol{\beta} - \mu)^2 \tag{2}$$

$\mathbf{M}_{i,\bullet}$ denotes here the i-th row of $\mathbf{M}$ representing the genomic data of individual i. The function which is minimized here is called the sum of squares. The solution to the minimization problem of Eq. (2) is given by the well-known OLS estimate

$$\begin{pmatrix}\hat{\mu}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}_{\text{OLS}} = \left(\begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix}^t \begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix}\right)^{-1} \begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix}^t \mathbf{y} \tag{3}$$

provided that the required inverse exists, which, in particular also means that n has to be greater than p.

In problems of statistical genetics, we often deal with a high number of loci and a relatively low number of observations. In this situation of $p+1 > n$, the solution to Eq. (2) is not unique but a vector subspace of which each point minimizes Eq. (2) to zero. Due to this overfit, the quality of predictions $\hat{\mathbf{y}}$ for genotypes which have not been used to estimate the parameters $(\hat{\mu},\hat{\boldsymbol{\beta}})$, are usually poor. An approach to overcome this problem is RRBLUP.

RRBLUP / GBLUP minimizes

$$\begin{pmatrix}\hat{\mu}\\ \hat{\boldsymbol{\beta}}\end{pmatrix}_{\text{RR}_\lambda} := \underset{(\mu,\boldsymbol{\beta})\in\mathbb{R}^{p+1}}{\arg\min} \sum_{i=1}^{n}(y_i - \mathbf{M}\boldsymbol{\beta} - \mu)^2 + \lambda \sum_{j=1}^{p}\beta_j^2 \tag{4}$$

for a penalty factor $\lambda > 0$. Using an approach of maximizing the density of the joint distribution of $(\mathbf{y},\boldsymbol{\beta})$, the model specifications of $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\beta^2)$ and $\boldsymbol{\epsilon}_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_{\boldsymbol{\epsilon}}^2)$ allow to determine the penalty factor as ratio of the variance components as $\lambda := \frac{\sigma_{\boldsymbol{\epsilon}}^2}{\sigma_\beta^2}$. We stress that Eq. (4) is not a pure ridge regression (RR), as the name RRBLUP might suggest, but a mixed model which treats μ and $\boldsymbol{\beta}$ differently by not penalizing the size of μ. This is the version, which is most frequently used in the context of genomic prediction (often with additional fixed effects).

The corresponding solution is given by

$$\begin{pmatrix}\hat{\mu}\\ \hat{\beta}\end{pmatrix}_{\mathrm{RR}_\lambda} = \left(\begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix}^t \begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix} + \lambda \begin{pmatrix}0 & \mathbf{0}_p^t\\ \mathbf{0}_p & \mathbf{I}_p\end{pmatrix}\right)^{-1} \begin{pmatrix}\mathbf{1}_n & \mathbf{M}\end{pmatrix}^t \mathbf{y}. \tag{5}$$

where $\mathbf{0}_p$ denotes the $p \times 1$ vector of zeros. The effect of the introduction of the penalization term $\lambda \sum_{j=1}^{p} \beta_j^2$ is that for the minimization of Eq. (4), we have a trade-off between fitting the data optimally and shrinking the square effects to 0. The method will only "decide" to increase the estimate $\hat{\beta}_j$, if the gain from improving the fit is greater than the penalized loss generated by the increase of $\hat{\beta}_j$.

First order epistasis: Polynomials of degree two

An extension of the additive model of Eq. (1) is a first order epistasis model given by a polynomial of degree two in the marker data [13–15]

$$y_i = \mathbf{1}_n\mu + \mathbf{M}_{i,\bullet}\boldsymbol{\beta} + \sum_{k=1}^{p}\sum_{j=k+1}^{p} h_{j,k}M_{i,j}M_{i,k} + \boldsymbol{\epsilon}, \tag{6}$$

where y_i is a polynomial with respect to the single variables $M_{i,j}$, $j = 1,\ldots,p$, and the product variables $M_{i,j}M_{i,k}$, $j,k = 1,\ldots,p$. In this context, components $M_{i,1}, M_{i,2}, \ldots, M_{i,p}$ are called monomials of degree 1, and components $M_{i,1}M_{i,1}, M_{i,1}M_{i,2}, \ldots, M_{i,1}M_{i,p}$ … $M_{i,2}M_{i,1}, M_{i,2}M_{i,2}, \ldots, M_{i,2}M_{i,p}$, …, $M_{i,p}M_{i,1}, M_{i,p}M_{i,2}, \ldots, M_{i,p}M_{i,p}$, are called monomials of degree 2. The same idea applies for higher interactions.

OLS Since the model is still linear in the coefficients, Eq. (3) represents the OLS solution, but with a modified matrix $\mathbf{M}$ including the products of markers as additional predictor variables.

eRRBLUP The extended RRBLUP is based on Eq. (6) and the assumptions of μ being fixed, $\beta_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\beta^2)$, $h_{j,k} \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_h^2)$ and $\epsilon_i \overset{i.i.d.}{\sim} \mathcal{N}(0,\sigma_\epsilon^2)$. In this case, the solution is also given by an analogue of Eq. (5), but with two different penalty factors, $\lambda_1 := \frac{\sigma_\epsilon^2}{\sigma_\beta^2}$ for additive effects and $\lambda_2 := \frac{\sigma_\epsilon^2}{\sigma_h^2}$ for interaction effects.

Translations of the marker coding

In quantitative genetics, column means are often subtracted from the original 0, 1, 2 coding of $\mathbf{M}$ to use $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}_n\mathbf{P}^t$ with $\mathbf{P}$ the vector of column means of $\mathbf{M}$ such that

$$\sum_{i=1}^{n} \tilde{M}_{i,j} = 0 \quad \forall j = 1,\ldots,p.$$

However, other types of translations, for instance a symmetric $\{-1,0,1\}$ coding or a genotype-frequency centered coding [16] can also be found in quantitative genetics' literature. Thus, the question occurs whether this has an impact on the estimates of the marker effects or on the prediction of phenotypes of genotypes which are not included in the training set.

The answer is that for the additive setup of Eq. (1), a shift from $\mathbf{M}$ to $\tilde{\mathbf{M}}$ will change $\hat{\mu}$ but not $\hat{\boldsymbol{\beta}}$ and any prediction $\hat{\mathbf{y}}$ will not be affected, neither for OLS, nor for RRBLUP (provided that λ is not changed). This invariance of the additive model does not hold for the extended RRBLUP method.

We give an example and discuss the effect of translations of the marker coding in a more general way afterwards.

Example 1 (Translations of the marker coding) *Let the marker data of five individuals with two markers be given:*

$$\mathbf{y} = (-0.72, 2.34, 0.08, -0.89, 0.86)^t \qquad \mathbf{M} = \begin{pmatrix} 2 & 2 \\ 1 & 2 \\ 2 & 0 \\ 2 & 1 \\ 1 & 0 \end{pmatrix}$$

Moreover, let us use the original matrix $\mathbf{M}$*, and the column mean centered matrix* $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}_5 \underbrace{(1.6, 1.0)}_{=:\mathbf{P}^t}$*. We consider the first order epistasis model*

$$y_i := \mu + \beta_1 M_{i,1} + \beta_2 M_{i,2} + h_{1,2} M_{i,1} M_{i,2} + \epsilon_i.$$

Then, we obtain the corresponding estimates based on i) an OLS regression, ii) a mixed model regression eRRBLUP-1 with $\lambda_1 = \lambda_2 = 1$*, and iii) a mixed model regression eRRBLUP-2 with* $\lambda_1 = 0$ *and* $\lambda_2 = 1$*, which penalizes only on the interaction term. The results are reported in Table 1.*

Table 1: Results from Example 1. "nc" denotes the use of the non-centered matrix $\mathbf{M}$ and "c" indicates the use of the centered matrix $\tilde{\mathbf{M}}$.

	OLS		eRRBLUP-1		eRRBLUP-2	
Estimates	nc	c	nc	c	nc	c
$\hat{\mu}$	1.83	0.33	1.81	0.33	2.69	0.33
$\hat{\beta}_1$	-0.97	-2.11	-0.89	-1.15	-1.54	-2.11
$\hat{\beta}_2$	1.88	0.06	0.71	0.09	1.03	0.11
$\hat{h}_{1,2}$	-1.14	-1.14	-0.48	-0.56	-0.57	-0.57
$\hat{\mathbf{y}}$						
	-0.91	-0.91	-0.47	-0.26	-0.61	-0.61
	2.34	2.34	1.38	1.45	2.07	2.07
	-0.11	-0.11	0.03	0.00	-0.39	-0.39
	-0.51	-0.51	-0.22	-0.13	-0.50	-0.50
	0.86	0.86	0.92	0.59	1.15	1.15

We summarize our observations from the reported results as follows:

- *Comparing the centered and uncentered versions of OLS, the estimated effects μ, β_1 and β_2 change, but the estimated interaction $\hat{h}_{1,2}$ as well as the prediction of* $\mathbf{y}$ *remain unchanged.*
- *Comparing the centered and uncentered versions of eRRBLUP-1, both codings give different estimates for all the parameters and the solutions produce different predictions for* $\mathbf{y}$.
- *Comparing the centered and uncentered versions of eRRBLUP-2, both codings give different estimates for μ, β_1 and β_2, but the same for $h_{1,2}$ and the same predictions for* $\mathbf{y}$.

The different cases presented in Example 1 have a certain systematic pattern, which we discuss in the following section.

Results

The observations made in Example 1 are explained by the following simple proposition, which has several interesting implications. For a mathematical proof of the statement, see the Appendix of this manuscript.

Proposition 1 *Let $\mathbf{M}_{i,\bullet}$ be the p vector of the marker values of individual i and let $f(\mathbf{M}_{i,\bullet}) : \mathbb{R}^p \to \mathbb{R}$ be a polynomial of degree D in the marker data. Moreover, let $\tilde{\mathbf{M}} := \mathbf{M} - \mathbf{1}_n\mathbf{P}^t$ be a translation of the marker coding and let us define a polynomial $\tilde{f}$ in the translated variables $\tilde{\mathbf{M}}$ by $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}) := f(\tilde{\mathbf{M}}_{i,\bullet} + \mathbf{P}^t) = f(\mathbf{M}_{i,\bullet})$. Then for any data $\mathbf{y}$, the sum of squared distances will be identical*

$$\sum_{i=1,\dots,n} (y_i - f(\mathbf{M}_{i,\bullet}))^2 = \sum_{i=1,\dots,n} (y_i - \tilde{f}(\tilde{\mathbf{M}}_{i,\bullet}))^2$$

and for any monomial m of degree D, the corresponding coefficient a_m of $f(\mathbf{M}_{i,\bullet})$ and $\tilde{a}_m$ of $\tilde{f}(\tilde{\mathbf{M}}_{i,\bullet})$ will be identical:

$$a_m = \tilde{a}_m.$$

Proposition 1 has the very simple statement that if we have a certain fit f based on $\mathbf{M}$, and we use the translated marker coding $\tilde{\mathbf{M}}$ in a second regression, the polynomial $\tilde{f}$ will fit the data with the same sum of squares and with the same predictions $\hat{\mathbf{y}}$ (due to the definition of $\tilde{f}$). Moreover, the coefficients of highest degree will be the same.

Since OLS is defined only by the minimal sum of squares, this also means that it is invariant to any translation of the coding, provided that the model structure allows the fit to adapt any f to the corresponding $\tilde{f}$ of Proposition 1. To allow this adaption, the possibility to adapt any coefficient of monomials of lower degree is required. We cannot adapt the regression completely if certain coefficients are forced to zero by the model structure. If a coefficient is equal to zero in f, it may be different from zero in $\tilde{f}$. We illustrate this with an example.

Example 2 (Models without certain terms of intermediate degree) *Let us consider the data* $\mathbf{M}$ *and* $\mathbf{y}$ *of Example 1 but with the assumption that marker* 2 *does not have an additive effect. Then*

$$\begin{pmatrix} \hat{\mu} \\ \hat{\beta}_1 \\ \hat{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 3.71 \\ -2.098 \\ -0.012 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} \tilde{\mu} \\ \tilde{\beta}_1 \\ \tilde{h}_{1,2} \end{pmatrix}_{OLS} = \begin{pmatrix} 0.334 \\ -2.11 \\ -1.162 \end{pmatrix}$$

and also the estimates $\hat{\mathbf{y}}$ *and* $\tilde{\mathbf{y}}$ *are different.*

Example 2 illustrates that “completeness" of the model is required to have the possibility to adapt to translations of the coding. More precisely, for any monomial of degree d, the model has to include all monomials of lower degree with these variables. If this is not the case, the adapted $\tilde{f}$ of Proposition 1 may not be a valid fit. Given that the model is “complete" in this sense, Proposition 1 has various implications. The following corollaries explain the results observed in our examples and highlight some additional properties of penalized regression methods in general. For all statements, it is assumed that penalty factors remain unchanged and that the model is complete in the sense that for any f, the corresponding $\tilde{f}$ is a valid fit.

Corollary 1 *For a model of any degree D, the OLS estimates of the coefficients of highest degree as well as the predictions* $\hat{\mathbf{y}}$ *are invariant with respect to translations of the marker coding.*

Corollary 1 is a result of the OLS method being defined only by the sum of squares, and f and the corresponding $\tilde{f}$ of Proposition 1 fitting the data with the same sum of squares when their respective coding is used. The fact that the coefficients of monomials of highest degree of f and $\tilde{f}$ are identical has been observed in Example 1.

Corollary 2 *For a model of degree D and a penalized regression which only penalizes the coefficients of degree D, the estimates of the coefficients of degree D as well as the predictions* $\hat{\mathbf{y}}$ *are invariant with respect to translations of the marker coding.*

Corollary 2 is a result of the following observation: for each f, its corresponding $\tilde{f}$ will have the same sum of squared distances and the same coefficients of highest degree (with the translated marker coding). Thus, it will have the same value for the target function of Eq. (4) which we aim to minimize. Because this is true for any polynomial f, it is in particular true for the solution minimizing the target function. A central point of Corollary 2 is that it is valid for any penalty on the size of the estimated coefficients of highest degree. The important requirement is that only the coefficients of highest degree are penalized.

Corollary 3 *In particular, predictions* $\hat{\mathbf{y}}$ *of RRBLUP are invariant with respect to translations of the marker coding, since we are dealing with a model of degree 1 and a regression that does not penalize the fixed effect (degree 0).*

Corollary 2 applied to complete models of degree 1 gives the result of Corollary 3, that is RRBLUP being invariant to translations of the marker coding. This fact has been previously proven using a marginal likelihood setup [17], or the mixed model equations [12].

Corollary 4 *An additive least absolute shrinkage and selection operator (LASSO) regression [18], ℓ_1 penalizing the marker effects but not the intercept is invariant to translations of the marker coding.*

Corollary 4 is a special case of Corollary 2.

Before we illustrate the impact of marker coding on estimated effect sizes with a publicly available data set, we give a small example, highlighting cases which are not invariant to translations of the marker coding. We recommend to use the data of Example 1 to validate the statements.

Example 3 (Regressions affected by marker coding)

a) Pure ridge regression of an additive model of Eq. 1 with a penalty on the size of μ is not invariant to translations.
b) RRBLUP without the fixed effect forced to zero is not invariant to translations of the marker coding.
c) An extended LASSO ℓ_1 penalizing additive effects and interactions is not invariant to translations of the coding.

Results on a wheat data set

To illustrate the practical implications, we compare the estimated interaction effects for different codings on a publicly available wheat data set.

Data We use the well investigated wheat data set published by Crossa *et al.* [19]. The data set includes the state of 1279 presence/absence markers of 599 genotyped wheat lines, and records on their yield performance in four different environments. The provided coding of the marker data is a $0,1$ coding. For more details on the data see Crossa *et al.* [19] or the `R` [20] package `BGLR` [21].

Calculating the interaction effects The assessment of the practical impact of translations of the marker coding on the effect estimates is difficult. Since in practice, the variance components and consequently the penalty factors are estimated on the data, the translations of the marker coding may have an additional indirect effect of changing the penalty factors. Moreover, there may be numerical issues changing the interaction effects when a large number of interactions is included. Thus, it seems difficult to separate these superposed effects. For this reason, we decided to restrict our model to the 100 most important markers and their interactions. In more detail, for each environment we perform RRBLUP, choose the 100 markers with highest absolute effect size and build a model with all pairwise interactions between them. The corresponding eRRBLUP includes the fixed effect μ, 100 additive effects and 4450 interactions. We prefer this approach to an approach of randomly selecting 100 markers, since approaches of restricting interactions to markers with large additive effects can be found in literature [22]. Moreover, we estimated the variance components only for the column mean centered coding and used the penalty factors also for the estimates with other codings. This is analogous to the fact that the translational invariance

of RRBLUP holds when the penalty factor remains fixed. For the estimation of the variance components, we used the `regress` package [23].

We compare three different codings: The originally provided $0, 1$ coding, a version translated by -0.5, that is a symmetric $-0.5, 0.5$ coding, and a coding in which the mean of each column is subtracted. We refer to these codings later as the *original* coding, the *symmetric* coding and *centered* coding.

For each of the environments, we compare the correlation of the estimated 4450 interaction effects for the three different codings. Table 2 illustrates that the estimated effect sizes are different when the coding is altered, but also that the correlation of the estimates is relatively high. In particular, the correlation of effect sizes for the original coding and the symmetric coding are at least on the level of 0.95 for all environments. The correlation of the estimates of each of these codings and the centered coding drops below 0.90 with a minimum of 0.80.

Table 2: Correlation of the estimates of the 4450 interactions with different marker coding. Colors indicate which data was used: Environment 1, 2, 3 or 4.

	$\hat{\mathbf{h}}_{symm}$		$\hat{\mathbf{h}}_{centered}$	
$\hat{\mathbf{h}}_{original}$	0.97	0.96	0.82	0.84
	0.95	0.95	0.80	0.83
$\hat{\mathbf{h}}_{symm}$		– –	0.86	0.87
		– –	0.85	0.88

Discussion

The illustrated problem of the coding having an impact on the estimates of interactions in penalized regressions is essential for quantitative genetics, where Hadamard products are often used to model interaction such as epistasis or gene by environment interaction [24]. The use of Hadamard products of covariance matrices to model interaction is an exact reformulation of the interaction effect model [14, 15]. In particular, this illustrates once more that the size of these interaction estimates should be interpreted with caution because a biological meaning is not necessarily given. This fact is also reflected by the fact that estimated variance components do not necessarily have a mechanistic meaning "genetic architectural" meaning [25].

The fact that the original and the symmetric coding produced estimates, which were more similar to each other than to the estimates derived from the centered coding also fits to previous observations, where the predictive ability of different codings was compared [12]. There, the column mean centered coding was found to be different from the codings in which each marker was coded in the same way. It was also found that the symmetric coding seems to outperform other codings slightly (with respect to predictive ability and in a cross-validation scheme).

It should be highlighted, that the problem does not seem to be a consequence of non-orthogonality of the predictor variables (marker values and their products), since these

problems would not appear in an OLS regression (provided that it exists), where the variables have the same coding and thus the same angle.

Finally, note that it has been reported that Gaussian Kernel regression [26] can be interpreted as a limit of the polynomial regression with increasing degree (and all possible monomials) [14]. Being a limit case of a method which is affected by translations of the coding, the question appears why the Gaussian Kernel regression is invariant to translations of the marker coding. It may be interesting to reconsider the limit behavior from a theoretical point of view.

Conclusion

We identified the cause of the coding-dependent performance of epistasis effects models. Our results were motivated by ridge regression, but do equally hold for many other types of penalized regressions, for instance for the ℓ_1 penalized LASSO. The fact that the estimated effect sizes depend on the coding highlights once more that estimated interaction effect sizes should be interpreted with caution in regard to their biological, mechanistic meaning. Moreover, this problematic of coding is not only present for marker by marker interaction, but for any mixed model in which interactions are modeled by Hadamard products of covariance matrices, in particular for gene by environment (G x E) models.

Appendix

Proposition 1 The fact that the goodness of fit remains the same, results from the definition of the polynomials. To see that the coefficients of monomials of highest degree are identical, choose a monomial $m(M_{l_1}, M_{l_2}, ..., M_{l_D})$ of the loci $l_1, ..., l_D$ of degree D of f. Multiplying the factors of $m(\tilde{M}_{l_1} + P_{l_1}, \tilde{M}_{l_2} + P_{l_2}, ..., \tilde{M}_{l_D} + P_{l_D})$ gives the same monomial $m(\tilde{M}_{l_1}, \tilde{M}_{l_2}, ..., \tilde{M}_{l_D})$ as a summand of highest degree, plus additional monomials of lower degree. Thus, the coefficients of monomials of degree D remain the same. □

Author's contribution

JWRM: Proposed to consider the topic; derived the theoretical results; JWRM and FR wrote the manuscript; All authors: Discussed the research

Acknowledgements

JWRM thanks KWS SAAT SE for financial support.

Author details
[1]KWS SAAT SE, Einbeck, Germany. [2]Universidad del Pacífico, Department of Finance, Lima, Perú. [3]University of Goettingen, Department of Animal Breeding and Genetics, Germany. [4]University of Goettingen, Chairs of Statistics and Econometrics, Germany. [5]University of Goettingen, Department of Animal Breeding and Genetics, Germany.

References

1. Meuwissen, T.H.E., Hayes, B.J., Goddard, M.E.: Prediction of total genetic value using genome-wide dense marker maps. Genetics **157**, 1819–1829 (2001)
2. Schaeffer, L.: Strategy for applying genome-wide selection in dairy cattle. J Anim Breed Genet **123**, 218–223 (2006). doi:10.1111/j.1439-0388.2006.00595.x
3. Habier, D., Fernando, R.L., Dekkers, J.C.M.: The impact of genetic relationship information on genome-assisted breeding values. Genetics **177**, 2389–2397 (2007). doi:10.1534/genetics.107.081190
4. Hayes, B.J., Visscher, P.M., Goddard, M.E.: Increased accuracy of artificial selection by using the realized relationship matrix. Genet Res **91**, 47–60 (2009). doi:10.1017/S0016672308009981
5. Hayes, B.J., Cogan, N.O.I., Pembleton, L.W., Goddard, M.E., Wang, J., Spangenberg, G.C., Forster, J.W.: Prospects for genomic selection in forage plant species. Plant Breeding **132**, 133–143 (2013). doi:10.1111/pbr.12037
6. Falconer, D.S., Mackay, T.F.C.: Introduction to Quantitative Genetics. Pearson Education, London, ??? (1996)
7. Schaeffer, L.: Application of random regression models in animal breeding. Livestock Production Science **86**(1), 35–45 (2004)
8. Henderson, C.R.: Best linear unbiased estimation and prediction under a selection model. Biometrics **31**, 423–447 (1975). doi:10.2307/2529430
9. Henderson, C.R., Quaas, R.L.: Multiple trait evaluation using relatives' records. J Anim Sci **43**, 1188–1197 (1976). doi:10.2527/jas1976.4361188x
10. Henderson, C.R.: Best linear unbiased prediction of breeding values not in the model for records. J Dairy Sci **60**, 783–787 (1977). doi:10.3168/jds.S0022-0302(77)83935-0
11. He, D., Parida, L.: Does encoding matter? a novel view on the quantitative genetic trait prediction problem. BMC bioinformatics **17**(9), 272 (2016)
12. Martini, J.W.R., Gao, N., Cardoso, D.F., Wimmer, V., Erbe, M., Cantet, R.J.C., Simianer, H.: 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 (2017). doi:10.1186/s12859-016-1439-1
13. Ober, U., Huang, W., Magwire, M., Schlather, M., Simianer, H., Mackay, T.F.C.: Accounting for genetic architecture improves sequence based genomic prediction for a *Drosophila* fitness trait. PLOS ONE **10**, 0126880 (2015). doi:10.1371/journal.pone.0126880
14. Jiang, Y., Reif, J.C.: Modeling epistasis in genomic selection. Genetics **201**, 759–768 (2015). doi:10.1534/genetics.115.177907
15. Martini, J.W.R., Wimmer, V., Erbe, M., Simianer, H.: Epistasis and covariance: how gene interaction translates into genomic relationship. Theor Appl Genet **129**, 963–976 (2016). doi:10.1007/s00122-016-2675-5
16. Álvarez-Castro, J.M., Carlborg, Ö.: A unified model for functional and statistical epistasis and its application in quantitative trait loci analysis. Genetics **176**(2), 1151–1167 (2007)
17. Strandén, I., Christensen, O.F.: Allele coding in genomic evaluation. Genetics Selection Evolution **43**(1), 25 (2011)

18. Tibshirani, R.: Regression shrinkage and selection via the lasso. Journal of the Royal Statistical Society. Series B (Methodological), 267–288 (1996)
19. 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.: Prediction of genetic values of quantitative traits in plant breeding using pedigree and molecular markers. Genetics **186**, 713–724 (2010). doi:10.1534/genetics.110.118521
20. R Core Team: R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria (2016)
21. de los Campos, G., Rodriguez, P.P.: BGLR: Bayesian Generalized Linear Regression. (2016). R package version 1.0.5. https://CRAN.R-project.org/package=BGLR
22. Kärkkäinen, H.P., Li, Z., Sillanpää, M.J.: An efficient genome-wide multilocus epistasis search. Genetics **201**(3), 865–870 (2015)
23. Clifford, D., McCullagh, P., Clifford, M.D.: The regress package. R package version, 1–3 (2014)
24. Pérez-Rodríguez, P., Crossa, J., Rutkoski, J., Poland, J., Singh, R., Legarra, A., Autrique, E., Campos, G.d.l., Burgueño, J., Dreisigacker, S.: Single-step genomic and pedigree genotype× environment interaction models for predicting wheat lines in international environments. The plant genome **10**(2) (2017)
25. Huang, W., Mackay, T.F.C.: The genetic architecture of quantitative traits cannot be inferred from variance component analysis. PLOS Genet **12**, 1006421 (2016). doi:10.1371/journal.pgen.1006421
26. Morota, G., Gianola, D.: Kernel-based whole-genome prediction of complex traits: a review. Front Genet **5**, 363 (2014). doi:10.3389/fgene.2014.00363